Symptom structures and individual patient profiles: novel insights into depression by the use of Nonmetric Multidimensional Scaling

Thesis (cumulative thesis)
Presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the Degree of Doctor of Philosophy

by Joël Bühler

Accepted in the Autumn Term 2013
on the Recommendation of the Doctoral Committee:
Prof. Dr. Damian Läge (main advisor)
Prof. Dr. Carolin Strobl

Zürich, 2013

Abstract

Attempts to measure depression via standardized rating scales reach back to the sixties. Since then, the two most commonly used depression rating scales to date, the Beck Depression inventory (Beck, Ward, Mendelsohn, Mock & Erbaugh, 1961) and the Hamilton Depression Rating Scale (Hamilton, 1960), have been analyzed with respect to their psychometric properties time and again. However, the results in the literature are diverging, as regards the suggested factor structure of the rating scales (Bagby, Ryder, Schuller, & Marshall, 2004; Brouwer, Meijer, & Zevalkink, 2013).

In the first part of the present thesis, the symptom structures of the two common rating scales, the revised Beck Depression Inventory (BDI-II; Beck, Steer & Brown, 1996) and the Hamilton Depression Rating scale (HAM-D), were analyzed by applying Nonmetric Multidimensional Scaling (NMDS) to contrast and compare the diverging factor analytic findings. In contrast to most conducted factor analyses of the BDI-II and the HAM-D, NMDS models the symptom-structure of the rating scales dimensionally instead of categorically. Furthermore, the symptom structures can be easily graphically represented as symptom spaces, if NMDS was conducted in a 2-dimensional space (which was the case in all the analyses within this thesis). Thus, 2-dimensional NMDS solutions excellently qualify to review the diverging factor structures found in the literature. The analyses of both rating scales revealed rather dimensional symptom structures which cannot be modeled adequately by traditional simple factor models. Thus, the diverging results in the literature could be attributed to generally insufficient models of the structures. Furthermore, the adequacy of these diverging factor models could be easily determined by the data's graphical representation in NMDS solutions.

In an additional study included in this thesis, a complex factor model of the BDI-II was derived from a previous NMDS solution. The model included an additional, activation related factor and it obtained better fit indices than the most reliable factor model to date, which is the G-factor model by Ward (2006). Thus, this study highlights the similarity between factor models and NMDS solutions and it exemplifies how NMDS solutions can be used to derive hypotheses about the data's underlying structure.

Although NMDS yields powerful capabilities to model symptom structures, it contains severe flaws with respect to standard error calculation: estimates of precision and stability can only be obtained by one single algorithm (Ramsay, 1977), which imposes additional assumptions upon the distribution of the data. There have been attempts to approach the problem of standard error calculation with bootstrap methods (Heiser & Meulman, 1983; Weinberg, Carroll, & Cohen, 1984), however, bootstrap procedures in NMDS differ with respect to the specific type of data. Moreover the procedures proposed in these studies were only insufficiently validated. To address the issue of uncertainty estimation, a simulation study was conducted in

the second part of the present thesis, which implemented a bootstrap procedure specifically tailored to rating scale data in an NMDS framework. The results suggested reasonably valid standard error regions in NMDS solutions when bootstrap methods were applied. Furthermore, an enhanced analysis method based on the bootstrap distributions of the NMDS solutions was shown to systematically reduce the bias in the data.

The third part of this thesis focused on individual symptom profiles of depressive patients. The profiles were analyzed with respect to two different aims. A first study analyzed the effect of symptomatically different subgroups of depression on response to treatment. Although depression is one of the most extensively researched disorders in psychology, its current definition is considered too broad and heterogeneous by many authors (e.g. (Baumeister & Parker, 2012; Carragher, Adamson, Bunting, & McCann, 2009; Fava et al., 1997; Lichtenberg & Belmaker, 2010). The concerns that depression indeed comprises different conditions are substantially driven by inconsistent findings in treatment response (Turner, Matthews, Linardatos, Tell, & Rosenthal, 2008). Accordingly, these authors suspect a lack of specific treatment strategies for the different subgroups. A great effort has been made to identify predictors of treatment response, yet the results have been humble: the findings in the literature are either fragmental or conflicting (Driessen & Hollon, 2010; Esposito & Goodnick, 2003; Hamilton & Dobson, 2002).

The study in this thesis followed a purely empirical approach to derive subgroups of depression: a latent Class Analysis was conducted in a large sample of depressive patients. The subgroups were then entered in a Linear Mixed Effects model to predict response to treatment. A significant effect of one subgroup emerged, which indicated slower response rates for its associated patients. The subgroup was identified as a melancholic subtype, however with reduced psychovegetative symptoms.

The last manuscript in this thesis explored the pitfalls and opportunities of NMDS in the analysis of individual patients' symptom profiles. Egli, Riedel, Möller, Strauss and Läge (2009) showed that NMDS analyses of patients' symptom profiles (patient maps) yielded promising results with respect to the separation of different diagnostic groups. However, the applicability of these patient maps in the diagnostic process depends on numerous factors that were not discussed by the authors. The last manuscript included in this thesis was written as an essay on these influencing factors. It extends and broadens the findings of (Egli et al., 2009) by exploring the systematic biases inherent in NMDS, when individual patients' symptom profiles are analyzed. Emphasis is laid on the issue of biased local inference, i.e. when similarities between profiles are inferred by the profiles' distances within small regions of the patient maps. Consecutively, an approach to optimize local inference is proposed.

Zusammenfassung

Versuche zur Quantifizierung depressiver Symptomatik durch standardisierte psychopathologische Inventare reichen zurück bis in die 60er Jahre. Bis heute werden insbesondere zwei dieser Inventare rege angewendet: Das Beck Depressionsinventar (Beck, Ward, Mendelsohn, Mock & Erbaugh, 1961) und die Hamilton Depressionsskala (Hamilton, 1960). Beide Instrumente wurden bezüglich ihrer psychometrischen Eigenschaften intensiv untersucht, wobei sich allerdings stark divergierende Befunde zeigten, insbesondere was die zugrundeliegende Faktorenstruktur anbelangt (z.B. Bagby et al., 2004; Brouwer et al., 2013).

Im ersten Teil dieser Dissertation wurde die Symptomstruktur der beiden Inventare mittels Nonmetrischer Multidimensionaler Skalierung (NMDS) untersucht. Denn im Gegensatz zu den meisten angewendeten Faktorenanalysen modelliert die NMDS die Symptomstrukturen der Inventare als dimensionale statt als kategoriale Strukturen. Darüber hinaus können NMDS Lösungen in einfacher grafischer Form dargestellt werden, sofern sich die Lösung auf zwei Dimensionen beschränkt, wodurch sie sich hervorragend für den Vergleich der faktorenanalytischen Resultate eignen. Beide Inventare zeigten in den NMDS Analysen eher dimensionale als kategoriale Strukturen, was die Modellierung der Daten mit den traditionell bevorzugten simplen faktorenanalytischen Modellen fehleranfällig macht. Die gefundenen dimensionalen Strukturen vermochten die divergierenden faktorenanalytischen Resultate in der Literatur zu erklären, und durch die grafische Repräsentation der Symptomstrukturen in den NMDS Lösungen konnte die Adäquanz der Faktorenmodelle eingeschätzt werden.

Eine weitere Studie dieser Dissertation befasste sich mit der Äquivalenz zwischen NMDS Lösungen und komplexen faktorenanalytischen Modellen. Darin wird ein Verfahren zur Ableitung von Faktorenmodellen aus NMDS Lösungen am Beispiel des BDI-II präsentiert, und es wird ein revidiertes Faktorenmodell für den BDI-II vorgeschlagen. Bezüglich Fit-Indizes übertrifft das vorgeschlagene Faktorenmodell, das einen zusätzlichen Aktivitätsfaktor beinhaltet, sogar das bislang führende Faktorenmodell zum BDI-II von Ward (2006).

Obwohl sich die NMDS zur Modellierung von Symptomstrukturen hervorragend eignet, besitzt sie einen tiefgreifenden Makel: Die Berechnung von Standardfehlern, und damit einhergehend die Darstellung von Konfidenzregionen, konnte bislang nur mit einem einzigen NMDS Algorithmus berechnet werden (Ramsay, 1977). Dieser stellt allerdings deutlich rigidere Voraussetzungen an die Daten, welche nur in Ausnahmefällen erfüllt sein dürften. Die Problematik fehlender Konfidenzregionen in NMDS Analysen wurde zwar schon früh erkannt, und unter Verwendung des Bootstrapverfahrens (Efron, 1979) wurde sogar eine Lösung des Problems vorgeschlagen (Heiser & Meulman, 1983; Weinberg et al., 1984). Allerdings gilt es zu beachten, dass sich das Verfahren je nach zugrundeliegender Datenbasis unterscheidet. Darüber hinaus wurde das Verfahren durch die genannten Autoren nur unzureichend validiert.

Um die Anwendbarkeit des Bootstrapverfahrens im Kontext der NMDS zu beurteilen, wurde eine Simulationsstudie durchgeführt, die speziell auf die Verwendung von Inventardaten abzielt und den zweiten Teil dieser Dissertation ausmacht. Die Studie zeigte eine vernünftige Übereinstimmung zwischen den durch das Bootstrapverfahren berechneten und den effektiven Konfidenzintervallen. Zusätzlich wird im Manuskript eine erweiterte Methode zur Berechnung von NMDS Lösungen präsentiert, welche die systematische Verzerrung der Daten reduziert.

Der dritte Teil der Dissertation beschäftigt sich mit den individuellen Symptomprofilen der Patienten, welche auf zwei unterschiedliche Ziele hin untersucht wurden. In einer ersten Studie wurde der Effekt von Subgruppen der Depression auf den Behandlungsverlauf analysiert. Denn obwohl die Depression in der psychologischen Forschung als eines der am besten erforschten Störungsbilder gilt, wird deren Homogenität von vielen Autoren angezweifelt (z.B. Baumeister & Parker, 2012; Carragher et al., 2009; Fava et al., 1997; Lichtenberg & Belmaker, 2010). Eine zu grobe Kategorisierung der Depression, und damit einhergehend das Fehlen von spezifischen Therapien, könnte demnach mitverantwortlich sein für die zum Teil mangelhaften Ansprechraten einzelner Patienten (Turner et al., 2008). Die Forschung zur Kategorisierung der Patienten in homogenere Subgruppen, die sich insbesondere durch unterschiedliche Ansprechraten auf die Behandlung auszeichnen sollten, blieb aber wenig aussagekräftig: Die Resultate sind entweder nur sehr eingeschränkt gültig oder sogar widersprüchlich (Driessen & Hollon, 2010; Esposito & Goodnick, 2003; Hamilton & Dobson, 2002).

In der vorliegenden Studie wurde ein statistisch-empirischer Ansatz zur Subgruppenbildung verfolgt: Durch den Einsatz der Latent Class Analyse wurden aus einem grossen Sample an depressiven Patienten Subgruppen gebildet und deren Effekt auf den Behandlungsverlauf durch eine nachfolgende Linear Mixed Effects Analyse überprüft. Eine der Subgruppen, die von ihrer Symptomatik her als Untergruppe der melancholischen Depression aufgefasst werden kann, zeigte einen deutlich geringeren Effekt der Behandlung bzw. längere Behandlungszeiten, was unterschiedliche Ansprechraten der Subgruppen auf die Therapie nahelegt.

Der letzte Artikel dieser Dissertation hatte zum Ziel, NMDS Analysen von individuellen Symptomprofilen (Patientenkarten) im Hinblick auf deren Anwendbarkeit im diagnostischen Prozess zu besprechen. Hinsichtlich Statusdiagnostik haben Egli et al. (2009) aufgezeigt, dass mittels NMDS Analyse eine Separierung unterschiedlicher Diagnosekategorien grundsätzlich möglich ist. Allerdings hängt die Strukturierung der Patienten in diesen Analysen von weiteren inhaltlichen und methodischen Faktoren ab, die in der Studie von Egli et al. (2009) nicht thematisiert werden konnten. Das letzte Manuskript, das im Stil eines Essays verfasst wurde, erweitert und verbreitert die Befunde von Egli et al. (2009) zu den methodischen Hürden der Patientenkarten und deren Einsatz in der Verlaufsdiagnostik. Ein Schwerpunkt des Artikels liegt auf dem methodischen Problem verzerrter Rückschlüsse aus den Patientenkarten, wofür bereits ein erster Lösungsansatz präsentiert wird.

Preface

The present thesis was inspired by the aim to create a tool that assists clinicians in day to day psychiatric care. The promising results by Läge, Egli, Riedel and Möller (2012) and Egli et al. (2009), as well as the findings from a study on diagnostic related groups funded by the Public Health Administration of Zurich that preceded this thesis, fostered the impression that applying Nonmetric Multidimensional Scaling (NMDS) to standardized psychopathological rating scales may substantially improve the workflow of psychiatrists in diagnostic and treatment decision making. After all, studies that examined the effect of feedback (to the clinician) on response to treatment in a mental health care setting generally found increased treatment efficacy when feedback was provided (e.g. Hatfield & Ogles, 2006; Lambert et al., 2005; Slade, Lambert, Harmon, Smart, & Bailey, 2008). Thus, in spring 2011, a project was launched, which has been devoted to turn NMDS analyzes of psychopathological rating scale data into a marketable product. Since summer 2011, the development of the theoretical foundations as well as the implementation of NMDS in a web based tool has been running under the project title PELION. The project was granted funding from the Commission for Technology and Innovation (CTI). Thus, the present dissertation was funded in large parts by the CTI within the scope of the larger project PELION.

This thesis was written as a cumulative dissertation. Thus, the included manuscripts differ in style (original studies / scientific report) and language (English / German) depending on the targeted place of publication.

Contents

Abstract II

Zusammenfassung IV

Preface VI

Contents VII

Introduction 9

NMDS: from psychophysics to clinical psychology 9
Symptom structures in NMDS analyses 10
Influence of the sample on the structure of an inventory 13
Patient structures in NMDS analyses 14
Included studies concerning symptom structures 15
Included study concerning methodological advancements in NMDS 18
Included studies concerning patient structures 19

Die Symptomstruktur des BDI-II: Kernsymptome und qualitative Facetten 22

Zusammenfassung 23
Abstract 23
Einleitung 24
Methodik 30
Ergebnisse 33
Diskussion 35

Activation as an overlooked factor in the BDI-II: a factor model based on core symptoms and qualitative aspects of depression 40

Abstract 41
Introduction 42
Methods 51
Results 52
Discussion 54
Acknowledgements 57

The Symptom Structure of the Hamilton Depression Scale (HAM-D): an NMDS analysis 58

Abstract 59
Introduction 60
Methods 63
Results 68
Discussion 71
Acknowledgements 77

Better bootstrap NMDS analyses – confidence regions and improved location estimates in Nonmetric Multidimensional Scaling 78

Abstract 79
Introduction 80
Methods 82
Analysis 1 85
Analysis 2 91
Discussion 95

The predictive power of subgroups: an empirical approach to identify depressive symptom patterns that predict response to treatment 101

Abstract 102
Introduction 103
Methods 105
Results 108
Discussion 112
Appendix 115

Berechnung und Interpretation von NMDS Patientenkarten für die Verlaufsdiagnostik: Erste Befunde 117

Zusammenfassung 118
Einleitung 119
Der Einfluss der Stresswert-Funktion und der Dimensionalität auf das Inverse-Interpretationsproblem in Patientenkarten 135
Diskussion 140

Conclusion 144

Future directions 146
Concluding remarks 147

References 149

Acknowledgements 161

Introduction

Methodical advancements in psychiatry are generally slow-going, especially so in the diagnostic process. Within the last thirty years, the indicators for the various mental disorders have not changed much and neither did treatment specificity. Even though the delineations of the disorders in the DSM occasionally shifted with its revision, the indicators of the disorders have prevailed since DSM-III: they consist of specific patterns of symptoms, along with a minimal duration of persistence. Thus, the DSM fostered empirical research of the diagnoses by quantifying the symptomatic underpinnings, but it also dissected the disorders into distinct categorical entities. A taxonomy consisting of a finite number of distinct categories may risk some major shortcomings though. One of these possible shortcomings is the discretization of intrinsically dimensional constructs (this matter was actually vigorously discussed for the revision of the DSM-V) and has been pointed out by many authors (e.g. Brown & Barlow, 2009; Egli, Riedel, Möller, Strauss, & Läge, 2009; Läge, Egli, Riedel, Strauss, & Möller, 2011). Another possible shortcoming is the heterogeneity of a disorder due to the pooling of different constituents as one single condition. In the field of depression research, heterogeneity is a frequently noted critique (e.g. Baumeister & Parker, 2012; Carragher, Adamson, Bunting, & McCann, 2009; Fava et al., 1997; Lichtenberg & Belmaker, 2010).

The current thesis is mainly concerned with the secondly noted shortcoming, i.e. heterogeneity, in depression. With its specific focus on the methods and by pushing the boundaries of commonly applied procedures to explore heterogeneity of depression, this thesis was able to add novel findings to the body of evidence in depression research. It thrives on the work by Egli et al. (2009), by Läge, Egli, Riedel and Möller (2012) and by Läge et al. (2011), who developed a framework to analyze data from psychopathological rating scales. Their work suggested that Nonmetric Multidimensional Scaling (NMDS) was indeed a valid method to examine the structure of disorders and psychiatric patients alike (Egli et al., 2009; Läge et al., 2012).

NMDS: from psychophysics to clinical psychology

Originally intended as a method to examine the psychological structure of perceived physical stimuli, Roger Shepard introduced NMDS to the methodical repertoire of psychophysicists (Shepard, 1962). He showed that only the rank order of similarity between stimuli suffice to map these stimuli into a low dimensional space. By applying NMDS, Shepard mapped the dissimilarities of different facial expressions for example, or the dissimilarities in the perception of colors, onto two dimensional spaces. The results were interpreted as psychological spaces, defined by arousal and pleasantness, and hue (the conventional color circle) for the facial and color stimuli respectively. The method of Multidimensional Scaling (MDS), i.e. the mapping of items described by dissimilarity data into a Euclidean Space, had already been proposed before

(Torgerson, 1958), however, it was not until Kruskal brought forth a sound mathematical foundation of NMDS that the method reached a widespread use (Kruskal, 1964). In the following years, a manifold of (N)MDS algorithms were developed implementing different approaches to minimize stress (the badness of fit criterion used in NMDS), different approaches to perform the nonmetric transformations of the similarity/dissimilarity data, and different approaches of weighting to correct for error in the data (e.g. De Leeuw, 1977; Kruskal, 1964; Läge, Daub, Bosia, Jäger, & Ryf, 2005; Ramsay, 1977).

Since the beginnings of NMDS, multidimensional scaling has expanded into many fields of psychological research such as market psychology (Carroll & Green, 1997), developmental psychology (Loeber & Schmaling, 1985) and neuropsychology (Abdi, Dunlop, & Williams, 2009). Furthermore, it has been applied in clinical psychology (Cohen, 2008; Läge et al., 2012; Steinmeyer & Möller, 1992) to examine the structure of psychopathological rating scales. The potency of NMDS in analyzing symptom structures was demonstrated by Läge et al. (2012), who reanalyzed the symptom structure of the AMDP inventory (Möller, 2009). They identified distinct regions in a two dimensional NMDS solution, which represented the underlying and repeatedly replicated factor structure of the AMDP. Moreover, they were able to identify an additional factor which had been postulated only theoretically at the time.

Applying NMDS did not stop at the analyses of symptom structures. In the contrary, Egli et al. (2009) turned the analysis of rating scales literally upside down and arrived at analyzing individual patients' symptom profiles (patient maps). Because the rating scale data of multiple patients are in the form of two-way two-mode data, NMDS analyses can focus on both ways, either on symptoms or on patients. The analysis of such patient structures, i.e. the similarity structure of a relatively small sample of patients, was an innovative approach that led to novel insights in the debate about the dimensionality of mental disorders: Egli et al. (2009) found different regions in the NMDS solution of schizophrenic, depressive and manic patients, each of which were predominantly populated with patients of one of the three groups. Furthermore, they noted a gradual transition between the group of depressive and schizophrenic patients, indicating gradual (instead of dichotomic) dissimilarities between these patients.

Symptom structures in NMDS analyses

NMDS analyses of psychopathological rating scales are conducted in two separate steps. In a first step, the similarity between the items is calculated by correlating them in a sample of completed rating scales (the individual symptom profiles). However, instead of a Pearson correlation, any other measure of coherence could be applied just as well. For example, if the symptom scores in the sample were markedly skewed, the nonmetric Spearman correlation coefficient may serve as a more adequate indicator of the similarity of the items. Also, distance measures

could be applied: for example the summed absolute differences between each pair of symptoms (note that in such a case, a large value indicates a small similarity between items). However, to integrate NMDS results in the previous findings in the literature, comparability of NMDS solutions with the results of conventional analyses methods (e.g. factor analyses) is aspired. Comparability is evidently highest if the methods share the same coefficient. Thus, the Pearson correlation coefficient is applied to compute NMDS solutions of symptom structures throughout this thesis.

In a second step, the similarities are transformed into distances (whereby small similarities become large distances) and mapped onto a low dimensional space. The procedure maintains the rank order of the similarities' reciprocals as good as possible in the mapped distances. The result is a low dimensional, Euclidean space in which each item has its specific location (and therewith fixed distances to any other item). The distances between the items reflect the structure of the similarities as good as possible within the bounds of (Euclidean) geometry in this lower dimensional space. Thus, highly similar items are located in close distance to each other, while less similar items are located farther apart. Generally, two dimensional spaces are already sufficient to model the symptom structure adequately. Two dimensional solutions can be depicted in a standard coordinate system and thus ease the interpretation of the results because the distances can easily be obtained visually. The term symptom maps will be used for these two dimensional NMDS solutions, as such a two dimensional, Euclidean space can be thought of as a geographic map: locations close to each other are generally much more alike than locations farther apart. In Figure 1, the symptom map of the Beck Depression Inventory II (BDI-II; Beck, Steer, & Brown, 1996) is exemplarily given to illustrate these two dimensional NMDS solutions.

It can be looked at NMDS solutions from two different perspectives. From a categorical perspective, one would focus on distinct regions in the solution space. A categorical solution is indicated by small, distinct clusters of items with large gaps between the clusters. These item clusters indicate coherent groups of items due to the high similarity of within-group items (close distances between the items of the same cluster) and the low similarity of between-group items (large distances between the clusters). The specific location of the items within the clusters is then considered as either random or irrelevant. Thus, an adequate factor model would best represent each of the item clusters as a separate factor. Contrarily, a dimensional perspective is indicated by soft transitions between the clusters. From a dimensional perspective, each items' specific location is considered yielding meaningful information with respect to the location of each other item. Thus, there is much more information when a dimensional perspective is taken. However, with the benefit of increased resolution comes the prize of lower stability: in each sample, the locations of the items are expected to vary (hopefully only slightly).

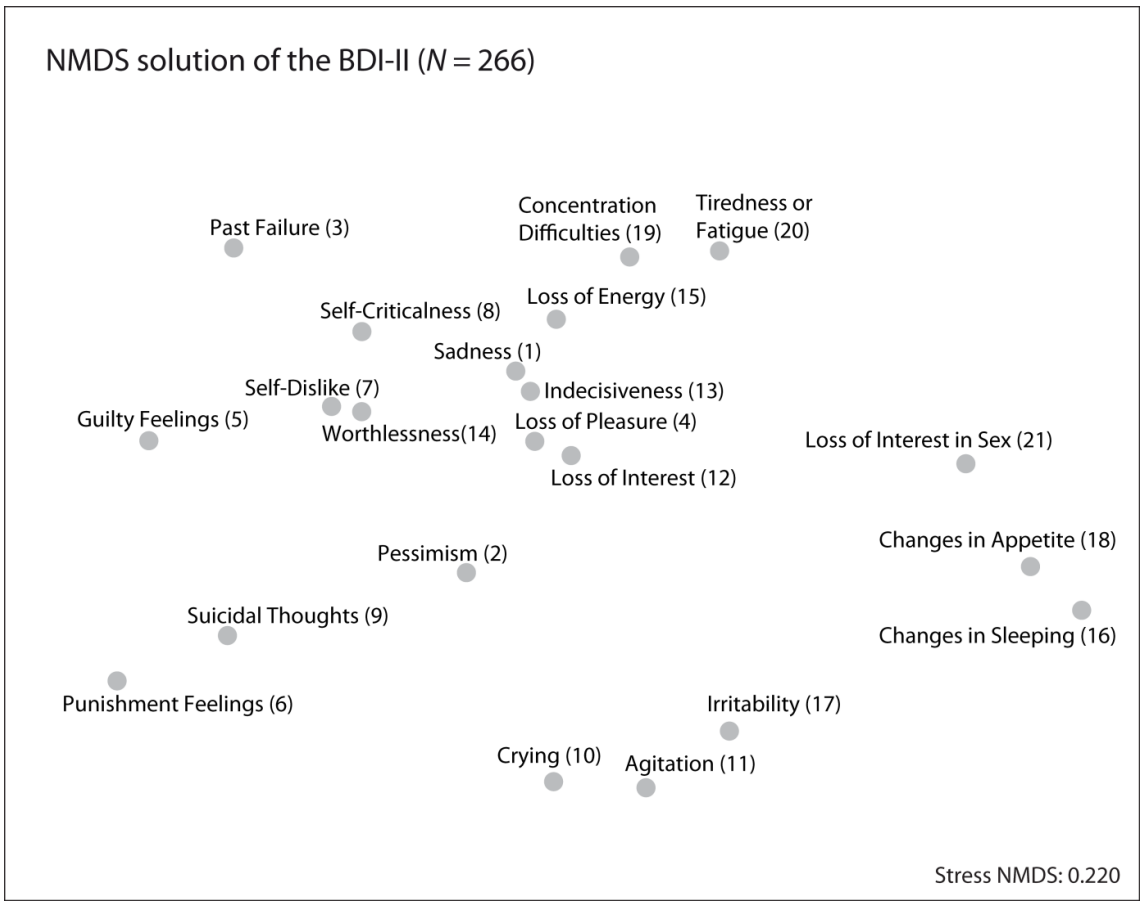

Figure 1. NMDS solution of the BDI-II. The distances between the items represent their pairwise similarities.

The easily obtainable dimensional information about symptom structures makes NMDS results not only an excellent framework to contrast and compare factor analytic findings of an inventory's item structure in the literature, but it represents also a superb foundation for the inference of complex factor models. Moreover, it allows an intuitive comprehension of the structure by the metaphorical power of the model: the complicated correlational structure of the symptoms is depicted as a map, where syndromes and symptom complexes are represented as distinct regions and the similarity of symptoms can quickly be inferred, simply by looking at their distances to each other. While the former argument is rather an argument for a scientific application of NMDS in the analysis of symptom profiles, the latter is first and foremost an argument for a beneficial applicability of the results in clinical practice.

Influence of the sample on the structure of an inventory

The symptom maps should not solely be regarded as representatives of the inventories' structures. Given certain preconditions, they allow a peek into the very structure of the respective disorder category. A first precondition requires a valid measure. Most of these inventories were developed decades ago; they were evaluated in many clinical studies and have been refined with respect to the items' wording and item selection. Thus, their validity in assessing the psychopathology of a disorder is undoubted. A second precondition requires that the analyzed sample comprises patients diagnosed with the disorder under study to establish a link between the structure of the inventory and the structure of the respective disorder. After all, an inventory's structure may well be sample dependent, as Reckase (2009) pointed out. Even though Reckase (2009) emphasized the measuring of cognitive skills and knowledge (and discussed it with respect to item response theory), his notes apply well to psychopathological inventories (and to factor analysis, which has mainly been applied to determine the structure of these inventories). The premises of his notions, which are that people must have different degrees of a skill (in the current case substitutable with severity of a disorder) and that the test (inventory) is able to assess these individual degrees, certainly also hold in the domain of psychopathology. Firstly, there is little doubt that psychopathological symptoms can be measured along a continuum of severity: most of the standardized psychopathological rating scales measure different degrees of symptom severity. Secondly, psychopathological inventories have been used since the 1960's to quantify the severity of mental disorders, which suggests clinical usefulness and appropriate criterion validity. With these premises met, the essence of Reckase's (2009) arguments shall be adapted to a depression inventory in the following paragraph to illustrate sample dependence of the inventories' structures.

Assume we collected depressive symptom data in two different samples: one sample consisted of patients with depression and one sample consisted of non-depressive students. When applying factor analysis to these two samples, a general factor "depression severity" could most likely be found in the depression sample (because the patients can actually be ordered on a continuum "depression severity"), but not in the student sample (simply because the students are not depressive and thus cannot be ordered on the continuum "depression severity"). Thus, systematic variation between the symptom profiles under study is an essential prerequisite for obtaining any factor structure. However, even if both samples yielded systematic variation in their symptom profiles, the variability must not necessarily be the same. Assume that some of the students were in the middle of an exam period; they may had slept badly which may had resulted in poor concentration and tiredness, and they may had felt stressed out which may had increased their irritability. Thus, instead of a factor "depression severity", a factor analysis of the student sample may obtain a factor "exam stress", which then again could not be found in the depressive sample. Accordingly, it should be kept in mind that any inventory not only has

one exact item structure, but that the structure is depending on both, the inventory and the analyzed sample instead.

However, given that an inventory is specifically designed (and able) to assess the symptoms of a disorder, and additionally, given that the symptoms themselves possess and can be measured on a continuum of severity, then the covariances between the symptoms in a sample of patients from the respective disorder can provide useful information about the structure of this same disorder. The variability of the patients' symptom scores, and the aggregation of the scores to different factors, may promote a more detailed understanding about the impairment of the patients' psychological functioning and, eventually, may provide additional, treatment relevant information.

Patient structures in NMDS analyses

The analysis of symptom profiles can, literally, be turned upside down whereupon similarities are calculated for patients (i.e. individual symptom profiles) instead of symptoms. The similarities between the patients' symptom profiles can be computed either as distance coefficients (in that case, the severity of the disorder has a strong influence on the resulting NMDS solution) or as correlation coefficients (in that case, the severity of the disorder has no influence on the resulting NMDS solution because of the preceding z-transformation). Analogously to symptom maps, NMDS analyses applied to patient data result in NMDS solutions, in which the patients' distances to each other represent the similarity structure between them. Thus, patients with highly similar symptom profiles are located in close distance to each other, while patients with dissimilar symptom profiles are located farther apart. Figure 2 shows an NMDS solution for 31 depressive patients based on distance coefficients of their BDI symptom profiles. Thirty patients were included with both, their symptom profiles at admission and at discharge. One patient (denoted as focus-patient in Figure 2) was included with all his weekly symptom profiles (*t0*-*t8*). As can be seen on Figure 2, the focus-patient responded well to his treatment: As the treatment proceeded, his symptom profiles increasingly resembled the other patients' symptom profiles at discharge (right hand side of Figure 2) and were located gradually further on the right hand side of Figure 2.

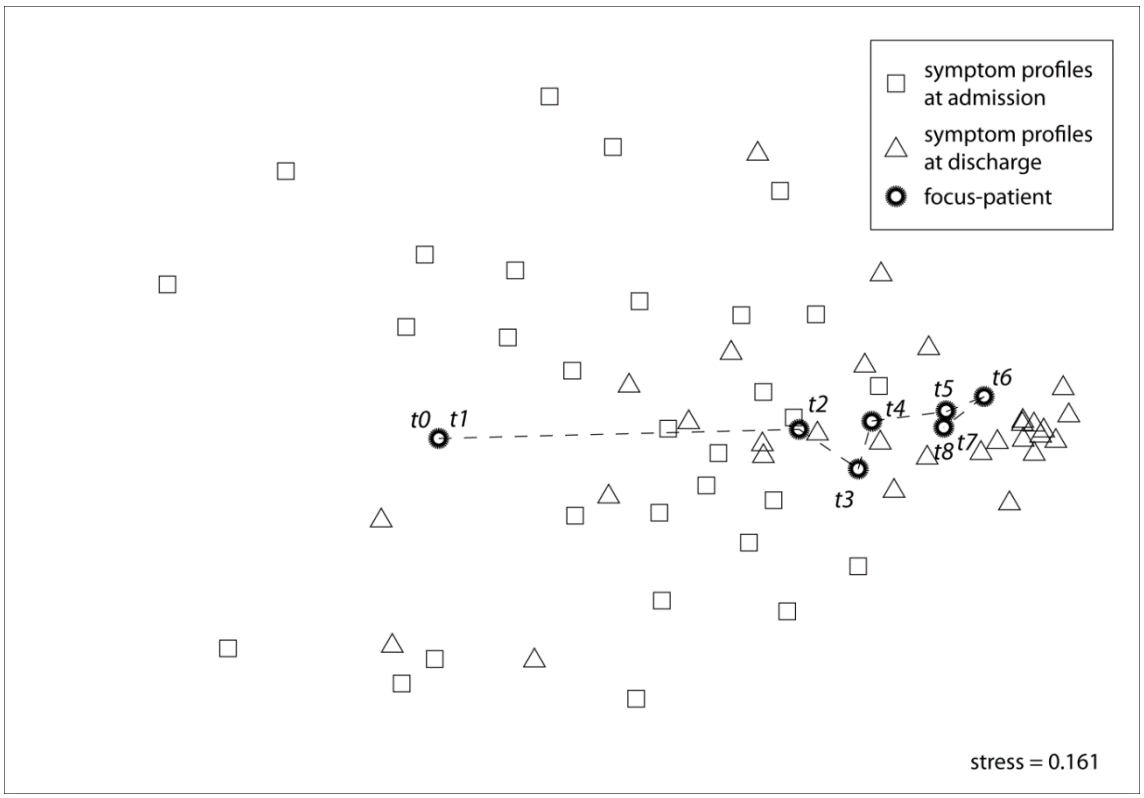

Figure 2. NMDS solution of depressive patients' similarity data. *t0 – t8* denote symptom profiles of one single patient at different time points.

Included studies concerning symptom structures

The first three studies of this thesis concern the symptom structure of the two most common rating scales in depression, a revision of the Beck Depression Inventory (Beck, Ward, Mendelsohn, Mock & Erbaugh, 1961) and the Hamilton Depression Rating Scale (Hamilton, 1960). While the manuscripts *"Die Symptomstruktur des BDI-II: Kernsymptome und qualitative Facetten"* and *„The Symptom Structure of the Hamilton Depression Scale (HAM-D): an NMDS analysis"* examined the symptom structure of the two rating scales via NMDS, the study entitled *"Activation as an overlooked factor in the BDI-II: a factor model based on core symptoms and qualitative aspects of depression"* thrived on the NMDS results of *"Die Symptomstruktur des BDI-II: Kernsymptome und qualitative Facetten"*, and it derived and examined a new factor model for the Beck Depression Inventory-II.

There has been an ongoing debate concerning the unidimensionality of both rating scales, the Hamilton Depression Rating Scale (HAM-D) and the BDI-II alike. Both scales have been widely accepted and their usefulness in measuring depression severity is unquestioned among clinicians. Thus, the debate on unidimensionality of the scales is not only of psychometric concern. Moreover, it may give rise to the question whether depression comprised a unidimensional construct altogether: after all, a dispute on the homogeneity of depression has been

going on for decades (e.g. Baumeister & Parker, 2012). Thus, understanding the symptom structure of these rating scales in a sample of depressive patients may additionally allow access to evidence about the structure of depression itself.

The study entitled *"Die Symptomstruktur des BDI-II: Kernsymptome und qualitative Facetten"* examined the symptom structure of the Beck Depression Inventory-II (BDI-II; Beck, Steer & Brown, 1996). The factor analytic results in the literature concerning the BDI-II are largely diverging. However, the main differences between the results can be condensed to two categories of solutions: simple factor models (mostly comprising two factors with either orthogonal or oblique factor structures e.g. Beck et al., 1996; Keller, Hautzinger, & Kühner, 2008) and complex factor models, which all implemented a bi-factor structure (e.g. Brouwer et al., 2013; Ward, 2006)[1]. Any bi-factor model consists of one general factor (which includes all items) and of at least two group factors, which do not overlap regarding their sets of associated items (Jennrich & Bentler, 2011). Hence, any item associated with a group factor indicates a complex factor structure. The most influential bi-factor model of the BDI-II to date includes a somatic and a cognitive group factor (Ward, 2006).

The two different model categories predicted diverging results in the NMDS analysis of the BDI-II. The simple factor models predicted distinct clusters of items, whereas the bi-factor models predicted soft transitions between the items. The study included in this thesis applied NMDS to the norming sample of the BDI-II German version (Hautzinger, Keller & Kühner, 2006) to obtain its symptom structures. The analysis revealed a pronounced dimensional symptom structure of the BDI-II, providing evidence for the superiority of bi-factor models. Additionally, the results indicated an activation related factor beside the postulated cognitive and somatic group factors.

The second study entitled *"Activation as an overlooked factor in the BDI-II: a factor model based on core symptoms and qualitative aspects of depression"* included in this thesis was conducted to replicate the results of the previous study *"Die Symptomstruktur des BDI-II: Kernsymptome und qualitative Facetten"*, which indicated a factor related to the activation level of the BDI-II symptoms, by applying factor analysis. Furthermore, the general principle to infer complex factor models from NMDS solutions was illustrated. Thus, a new factor model was derived from the NMDS solution of the BDI-II, and it was compared with the most reliable factor

[1] In a simple factor structure, each item is associated to maximally one factor. Thus, if two items are associated to the same factor, they are essentially equivalent. The difference in the factor loadings of these items is then axiomatically assumed to originate from different amounts of error in the measure. In contrast, items may be associated to more than one factor in a complex factor structure. Complex factor structures allow items to differ from one another even if they load on the exact same factors: the items values may differ due to different linear combinations of the associated factors and thus yield different meanings despite their equivalent loading pattern.

model in the literature to date, which is the model by Ward (Brouwer et al., 2013; Quilty, Zhang, & Bagby, 2010; Ward, 2006).

The deduction of the BDI-II factor model was guided by the individual locations of the items in the NMDS solution of the BDI-II. It was hypothesized that the symptoms' locations were defined by linear combinations of three factors: a cognitive, a somatic and an activation related factor. Even though activation has been frequently identified as an important categorization criterion in the history of depression research (e.g. Koukopoulos & Koukopoulos, 1999; Shorter, 2007), the factor model in the manuscript was the first to postulate an activation factor in the BDI-II.

The results indicated a good approximation of the postulated factor model to the data. Moreover, the loading patterns of the items on the factors were concordant in the main with the hypothesized linear combinations. There were some unexpected results worth noting though. Firstly, the activation factor was mainly defined by items with a low level of activation, which undermined to some extent the hypothesis that the suggested items defined two ends of a bipolar scale. Furthermore, two items, "crying" and "irritability" did not reveal a significant loading on the activation factor, despite their hypothesized association with a high level of activation. Nevertheless, the proposed factor model surpassed the group factor model by Ward with respect to the conventional fit indices (Ward, 2006).

The third study entitled *„The Symptom Structure of the Hamilton Depression Scale (HAM-D): an NMDS analysis"* was conducted to examine the symptom structure of the HAM-D (Hamilton, 1960), more specifically, the revised German version of the HAM-D (Collegium Internationale Psychiatriae Scalarum, 1977). Similar to the results of the BDI-II, the findings on the factor structure of the HAM-D in the literature are diverging. However, in contrast to the BDI-II, the models could not be easily allocated to a small number of distinct categories. Instead, the models mainly differed in the number of factors assumed, which varied between 2 (Steinmeyer & Möller, 1992) and 8 (Giesen, Bäcker, & Hefter, 2001; O'Brien & Glaudin, 1988). Similarly to the approach in *"Die Symptomstruktur des BDI-II: Kernsymptome und qualitative Facetten"*, an NMDS analysis of the HAM-D items based on depressive patients' symptom profiles was conducted to explore the symptom structure of the HAM-D. Furthermore, a meta-analytic strategy described by Loeber and Schmaling (1985) and Frick et al. (1993) was applied, which enabled analyzing the results of the factor analyses in the literature directly. The two NMDS solutions, which were calculated from independent data sets, were then compared with respect to their symptom structure.

The two analyses revealed very similar symptom structures: in both analyses, rather dimensional structures emerged and the locations of most items were in good accordance with each other as well as with theoretical considerations. Coherent groups of symptoms could be

identified and were labeled insomniac, gastrointestinal, somatic, psychotic, and core symptoms. Furthermore, a group of symptoms with increased activation was identified.

Included study concerning methodological advancements in NMDS

The studies presented in the previous section applied NMDS to examine the symptom structure of the BDI-II and the HAM-D and the results were replicated by either using a different methodological approach (a factor analysis to replicate the structure of the BDI-II) or different data (the meta-analytic procedure to replicate the structure of the HAM-D) respectively. Even though the replications revealed good overall concordance between the structures, the stability and range of variation of the symptoms' locations in the NMDS solutions could not be estimated.

Thus, computationally derived confidence regions (i.e. confidence intervals in more than one dimension) may greatly benefit the interpretability of NMDS solutions. While for many statistical models confidence intervals of their estimates can be obtained directly as derivatives of an explicitly defined distribution function, most NMDS algorithms make no such explicit assumptions about the modeled data. However, methods exist to compute confidence intervals without assuming any specific probabilistic model. Probably the most common of these methods is the bootstrap (Efron, 1979; Efron & Tibshirani, 1986) which was suggested to be applied to compute confidence regions in NMDS (Heiser & Meulman, 1983; Weinberg et al., 1984) about thirty years ago. However, assumingly due to the lack of computational power, the procedure had not been systematically evaluated with respect to the precision of the estimated confidence intervals. Thus before generally applying the bootstrap in NMDS, a systematic evaluation of the method was of cardinal importance.

The analysis of psychopathological data, as suggested by Egli et al. (2009) and Läge et al. (2012), differs from the classical application of NMDS. Instead of direct similarity ratings, they applied correlation coefficients and distance measures, which were derived from two-way two-mode data. Although those indirect similarity/dissimilarity (hereinafter proximity is used to refer to both coefficient types) coefficients have become a common measure to apply NMDS to (Borg & Groenen, 2005), they lack the direct proximity's quality of independence of each other and, as was shown in the study, introduce a bias by aggregating the data.

The fourth study of this thesis approached both presented issues and was entitled "*Better bootstrap NMDS analyses – confidence regions and improved location estimates in Nonmetric Multidimensional Scaling*". Firstly, applicability of the bootstrap to calculate confidence regions in NMDS was thoroughly evaluated by a large simulation study, and secondly, an extended NMDS procedure was proposed to reduce the bias in the calculation of proximity coefficients. The results indicated good applicability of the bootstrap in NMDS, as long as the percentile method was applied to estimate the confidence intervals: the more sophisticated methods of

bias correction in fact worsened the validity of the method. It was argued in the study that these results followed from two sources of error: a bias introduced by the calculation of proximity and the issue of local minima, a problem which has been extensively discussed in the literature (Groenen & Heiser, 1996). Concerning the reduction of the bias in proximity computation, an improvement in the mean error of the distance estimates was found when the extended procedure was applied instead of the traditional NMDS analysis.

Included studies concerning patient structures

The fifth study in this thesis examined the structure of HAM-D profiles from individual, depressive inpatients at admission. It was entitled *"The predictive power of subgroups: an empirical approach to identify depressive symptom patterns that predict response to treatment"*. The study was specifically designed to examine the heterogeneity of the patients' symptom profiles and to test the predictive power of subgroups with respect to treatment response. Provided that depression comprises a number of different conditions, as has been suggested by many authors (Baumeister & Parker, 2012; Carragher et al., 2009; Fava et al., 1997; Lichtenberg & Belmaker, 2010), patients affected by different conditions were expected to exhibit distinct symptom patterns. Vice versa, patients affected by the same conditions should exhibit similar symptom patterns. Based on this premise, there have already been attempts to identify homogenous subgroups of depressive patients with various statistical classification techniques (e.g. Aggen, Neale, & Kendler, 2005; Blazer et al., 1989; Carragher et al., 2009; Cox, Enns, & Larsen, 2001). However, the empirically derived classification attempts did not have a large impact on the theoretically coined discussion of depression heterogeneity. In contrary, those attempts were even appraised as unsuccessful by some authors (e.g. Lichtenberg & Belmaker, 2010). One explanation of this lack of impact may be that the purely empirically driven approach neither substantially added new knowledge to the already theoretically postulated subgroups of depression (generally, the subgroups were similar to the theoretically derived ones) nor proved meaningful regarding treatment outcome (the dependency of empirically derived subgroups on treatment response was simply ignored).

In the study included in this thesis, the subgroups were obtained by applying Latent Class Analysis (LCA) to the HAM-D data of depressive inpatients at admission. One common shortcoming of the existing LCA studies is that none of the studies applied centering of the symptom data prior to the main analysis. Thus, the inherent quality of depression (i.e. the specific pattern of symptoms) was necessarily intertwined with depression severity (i.e. the symptom score). Although depression severity may be a reasonable grouping criterion in some situations, it probably does not qualify to distinguish different depressive conditions (which may all exhibit lesser or more severe states). Therefore, a reanalysis of symptom profile data applying

LCA to previously centered data seemed promising. The classes obtained from the LCA were considered to comprise subsets of patients, which were all affected by the same condition.

Consecutively, the effect of different subgroups of depression on treatment response was assessed with a Linear Mixed Effects (LME) model. In the LME model, the classes obtained in the LCA were used to predict response to treatment, which was measured as the change in the patients' summed HAM-D_{17} scores during their treatment. Five classes were obtained from the LCA, all of which showed substantially different symptom patterns. Four of the five classes were identified as relating to the frequently postulated subgroups melancholic and anxious depression (Baumeister & Parker, 2012). However, LCA divided each of the two subgroups again in two different categories. The fifth class was characterized by substantially elevated scores on the item suicide and substantially decreased scores on the anxiety related items psychic and somatic anxiety. In the following LME analysis, the one group of patients most similar to the pattern of melancholic depression revealed a significantly slower response to treatment compared to all other subgroups. Patients with melancholic depression have been found to reveal slower response to treatment before (e.g. Fava et al., 1997); however, other researchers have found no significant effect of melancholic depression on treatment response (Fournier, DeRubeis, Shelton, Hollon, Amsterdam, & Gallop, 2009; Jarrett, Minhajuddin, Kangas, Friedman, Callan, & Thase, 2013). It was hypothesized in the present study that melancholic depression may be comprised by two different conditions, of which only one shows substantially slower response to treatment.

The sixth manuscript entitled *"Berechnung und Interpretation von NMDS Patientenkarten für die Verlaufsdiagnostik: Erste Befunde"* was written as an essay on the applicability of patient maps in clinical practice, in particular on the applicability of NMDS to analyze individual patients' symptom profiles. Applicability was discussed with respect to comparing patients based on their symptom profile similarities, providing feedback during the course of treatment and thus, ultimately, fostering evidence based medicine in psychiatric and psychological care.

It has been shown that patient maps reveal distinct regions that concur with the patients' diagnoses (Egli et al., 2009). However, a general applicability of patient maps in clinical practice, especially with respect to the course of treatment, cannot be inferred from the study by Egli et al. (2009). Firstly, Egli et al. (2009) only examined patient maps with respect to their diagnostic properties at admission. Thus, the properties of patient maps with respect to the assessment of symptomatic change during treatment still remained in the dark. Secondly, the benefit of patient maps in the selection of treatment regimen heavily rely on the assumption that meaningful patient structures emerge with respect to differential treatment outcome. Thirdly, inference of treatment regimens for individual patients rests upon a local interpretation of specific regions within the patient maps (i.e. the region about a focus-patient for whom inferences shall

be drawn). However, since NMDS solutions are not optimized with respect to local structures (the regions about a focus-patient) but instead are optimized with respect to the global structure of all patients included in the patient space, local interpretations may be biased.

The manuscript discussed the application of patient maps in the domain of psychiatric treatment planning and evaluation. Mainly, the manuscript focuses on the current methodological and interpretative hurdles and pitfalls in such an application. It was argued that the interpretation of the similarity of symptom profiles within small regions of the patient space (which arises from the interest in the focus-patient) is substantially biased. Because only small distances are considered in the interpretation, errors occur mainly by overestimating the similarities between patients (false-positive bias). A minor revision in the NMDS algorithm was proposed to attenuate this false-positive bias and a first experiment yielded encouraging results regarding the performance of the revision. Furthermore, the potential limitations to validly assess profile differences in low dimensional spaces were highlighted, especially as regards multidimensional psychopathological inventories.

Die Symptomstruktur des BDI-II: Kernsymptome und qualitative Facetten

[The symptom structure of the BDI-II: core symptoms and qualitative facets]

Joël Bühler[1], Ferdinand Keller[2], Damian Läge[1]

[1]Universität Zürich, Schweiz, Psychologisches Institut, Angewandte Kognitionspsychologie
[2]Universitätsklinikum Ulm, Klinik für Kinder- und Jugendpsychiatrie/Psychotherapie

Submission status:

Published in *Zeitschrift für Klinische Psychologie und Psychotherapie, 41*(4), 2012, 231-242.

Authors' contributions:

Joël Bühler: Review of the literature, execution of the analyses, interpretation of the results, writing of the manuscript

Ferdinand Keller: Provision of the data, revision of the manuscript, revision of the wording in the NMDS solution

Damian Läge: Development of the research question, supervision and discussion of Joël Bühler's contributions, revision of the manuscript

Zusammenfassung

Theoretischer Hintergrund: Das BDI-II deckt mit 21 Depressionssymptomen die Breite der Depression ab und ist geeignet, die Symptomstruktur der Depression zu erforschen. Eine Vielzahl an faktorenanalytischen Studien erbrachte aber bislang keinen Konsens. **Fragestellung:** Nonmetrische Multidimensionale Skalierung (NMDS) soll daher die divergierenden Resultate einordnen und ein Modell der Symptomstruktur schaffen. **Methode:** Mittels NMDS wird die Symptomstruktur des BDI-II (N=266 Depressive) in einen 2-dimensionalen Raum abgebildet. **Ergebnisse:** Die NMDS-Lösung legt eine Facettenstruktur nahe, welche von den faktorenanalytischen Modellen bislang unzulänglich erfasst wird. **Schlussfolgerungen:** Neben einem generellen Kernsyndrom finden sich fünf spezifische Facetten (verminderte Aktivierung, psychovegetative Symptome, gesteigerte Aktivierung, Hoffnungslosigkeit und negative Einstellung zum Selbst), die die Heterogenität der Symptomatik innerhalb der Depression aufzeigen und damit die Existenz von Subtypen in der Depression nahelegen.

Schlüsselwörter: Beck Depressionsinventar II; Depression; Nonmetrische Multidimensionale Skalierung

Abstract

Background: The BDI-II (Beck Depression Inventory-II) is a commonly used self-assessment scale for depression and consists of 21 symptoms. To investigate the symptom-structure of depression, the BDI-II has been analyzed by many factor analytic studies; results, however, are inconsistent. **Objective:** To reexamine the diverging results on the basis of Nonmetric Multidimensional Scaling (NMDS). **Methods:** NMDS was applied to BDI-II data of 266 depressed patients. **Results:** A facet oriented symptom-structure was obtained, which cannot be captured adequately by the existing factor models. **Conclusions:** The NMDS solution reveals a depressive core-syndrome and five specific facets (diminished arousal, psychovegetative symptoms, increased arousal, hopelessness, and negative attitude towards self), which indicate structural heterogeneity in depressive symptoms suggesting the existence of subtypes within depression.

Key words: Beck Depression Inventory II; Depression; Nonmetric Multidimensional Scaling

Einleitung

Im Vorlauf zu den geplanten Revisionen des DSM-IV und der ICD-10 ist seit einigen Jahren die Diskussion zur Homogenität der Diagnosekategorie „Major Depression“ erneut aufgeflammt. Die weitergehende Unterteilung der Depression in spezifischere Subgruppen und deren Verankerung in den Manualen wird vielerorts gefordert (z.B. Damm, Eser, Schüle, Möller, Rupprecht & Baghai, 2009; Fink & Taylor, 2007; Joiner, Walker, Pettit, Perez & Cukrowicz, 2005; Parker, 2007; Shorter, 2007; Stewart, McGrath, Quitkin & Klein, 2007). Geeignete Grundlage für die Subklassifikation der Depression und für die entsprechende Klassifikation der depressiven Patienten wäre ein Wissen über die Struktur der Symptome. Eine allgemein anerkannte Liste dieser Symptome liegt beispielsweise mit dem DSM vor und wird von Tests wie dem BDI erfasst (Beck, Ward, Mendelsohn, Mock & Erbaugh, 1961), einem weit verbreiteten Inventar zur Erfassung des Schweregrades der Depression.

Seit 1996 liegt mit dem BDI-II (Beck, Steer & Brown, 1996) eine revidierte Fassung des Inventars in englischer und seit 2006 auch in deutscher Sprache vor (Hautzinger, Keller & Kühner, 2006). Die Revision zollte Kritiken unter anderem von Moran und Lambert (1983) Rechnung, dass die diagnostischen Kriterien des – damals aktuellen – DSM-III nicht vollständig im Fragebogen abgebildet würden. Das BDI ist kein Test mit auf Messwiederholungen hin konstruierten Items, sondern es will die tatsächlichen Symptome der Depression einzeln für sich erfassen. Die Items sind deswegen mit den jeweiligen vier Aussagen auch so strukturiert, dass der Schweregrad jedes Items für sich genommen möglichst gut eingeschätzt werden kann. In der Summe ergibt sich daraus zwar der nützliche Gesamtschweregrad der Erkrankung, doch ist der Test nicht auf einige wenige Subskalen (mit Untersummen) hin konstruiert worden. Vielmehr entspricht die gemessene Varianz der „natürlichen“ Varianz der Symptome bei depressiven Personen, eben weil die Symptomliste den gesamten Bereich der Depression nach DSM-Kriterien abbildet.

Mit dem Schritt vom DSM-III zum DSM-IV wurde deswegen auch eine Revision des BDI notwendig, und mit ihr auch eine wiederholte Prüfung der psychometrischen Eigenschaften des Fragebogens. Inzwischen liegt hierzu eine Vielzahl an internationalen Studien vor, mit einem methodischen Fokus auf der Faktorenanalyse. Diese Studien suchen nach einer Ebene zwischen dem Summenscore (welcher aufgrund der grundsätzlich positiven Korreliertheit aller 21 Symptome als Messung des Schweregrades sinnvoll ist) und den Einzelsymptomen. Diese gesuchte Zwischenebene kann allein schon aus Gründen der kognitiven Vereinfachung für sinnvoll erachtet werden (um die 21 Symptome zu einer übersichtlichen Anzahl kleinerer Pakete zu bündeln), sie mag aber auch für die Klassifikation von Subtypen der Depression für viele Forscher Berechtigung besitzen.

Der bislang beschrittene faktorenanalytische Weg, zu den gesuchten Zwischenebenen zu gelangen, führte zu einer kategorialen Differenzierung „kognitiv vs. somatisch“, darüber hinaus jedoch zu keinem einheitlichen Ergebnis; im Gegenteil, die Resultate zur Faktorenstruktur im BDI-II divergieren erheblich, z.T. sogar innerhalb derselben Publikation: So fanden Beck et al. (1996) mittels exploratorischer Faktorenanalyse bei einer Stichprobe aus 500 ambulanten Patienten zwei oblique Faktoren, welche sie als kognitiv und somatisch-affektiv identifizierten. Auf den kognitiven Faktor luden die 9 Items „Versagensgefühle“, „Wertlosigkeit“, „Traurigkeit“, „Pessimismus“, „Schuldgefühle“, „Bestrafungsgefühle“, „Selbstablehnung“, „Selbstkritik“ und „Suizidgedanken“, auf dem somatisch-affektiven Faktor die übrigen 12 Items. In derselben Studie wurden auch 120 Studenten untersucht (Beck et al., 1996). In der studentischen Stichprobe zeigten sich zwar auch wieder 2 oblique Faktoren, diesmal allerdings als somatischer Faktor (5 Items) und kognitiv-affektiver Faktor (16 Items). Der somatische Faktor wurde hier begründet durch die Items „Energieverlust“, „Ermüdung“, „Schlaf“, „Appetitveränderung“ und „Konzentrationsschwierigkeiten“, der kognitiv-affektive durch die Übrigen. Diese beiden Zuordnungsmuster wurden in den meisten Studien repliziert, die Gruppe der affektiven Items blieb dabei immer etwas unsicher in der Zuordnung: Je nach Stichprobe und Studie fielen sie dem somatischen oder dem kognitiven Faktor zu.

Aufgrund der durchgehenden positiven Korreliertheit aller 21 Symptome legten Arnau, Meagher, Norris und Bramson (2001) einen anderen Ansatz zur Extraktion der Faktoren im BDI-II vor: Sie wendeten die Schmid-Leiman Transformation an (Schmid & Leiman, 1957), bei der Faktoren 2. Ordnung über die Variablen (statt über die Faktoren) extrahiert und die gemeinsame Varianz aus den Faktoren 1. Ordnung herauspartialisiert werden. Dies zeigte einen generellen Faktor (G-Faktor) 2. Ordnung, der die meiste Varianz der beiden Faktoren 1. Ordnung (somatisch-affektiv, kognitiv) erklären konnte. Obwohl die resultierenden Faktoren 1. Ordnung z.T. noch substantielle Ladungen aufwiesen, wurde die Lösung mangels Signifikanz zurückgewiesen. Die vorgeschlagene Faktorenstruktur eines G-Faktors und zweier davon unabhängigen spezifischen Faktoren wurde von Ward (2006) wieder aufgegriffen. Die Zuordnung der Items zu den Faktoren nahm er theoriegeleitet vor, wobei alle Items von einem G-Faktor beeinflusst sind und nur ein reduziertes Set an Items als Indikatoren für den kognitiven respektive den somatischen Faktor dienen sollten. Die Symptome mit spezifischer kognitiver oder somatischer Varianz würden so von den unspezifischen affektiven Items getrennt, welche ausschließlich vom G-Faktor beeinflusst und entsprechend in der Zuordnung in einem Zwei-Faktoren-Modell unsicher wären. In einer Reanalyse von fünf Datensätzen des BDI-II (Beck et al., 1996; Buckley, Parker & Heggie, 2001; Steer, Ball, Ranieri & Beck, 1999; Steer & Clark, 1997; Whisman, Perez & Ramel, 2000) zeigte er mittels konfirmatorischer Faktorenanalyse eine gute Anpassungsleistung seines Modells an die Daten.

Tabelle 1 fasst die Resultate der exploratorischen Faktorenanalysen und des konfirmatorisch getesteten Modells von Ward (2006) zusammen. Die fett gedruckten Symptomnummern

sind diejenigen, die bei allen Modellen entweder dem kognitiven oder dem somatischen Faktor zugeordnet wurden (kleinster gemeinsamer Nenner). Einzig die Studie von Osman, Downs, Barrios, Kopper, Gutierrez und Chiros (1997), welche eine substantiell andere Faktorenstruktur als die übrigen Studien zeigt, wurde bei der Eruierung des gemeinsamen Nenners nicht mit einbezogen. (Damit wäre eine weitere Reduktion der „somatischen Symptome“ von fünf (15, 16, 18, 19, 20) auf nur gerade zwei (16, 18) einhergegangen, was angesichts der gesamthaften Datenlage nicht mehr sachgerecht erscheint.)

Die unterschiedlichen Modelle wurden in der Folge in unterschiedlichen Studien mittels konfirmatorischer Faktorenanalyse überprüft. Eine der umfangreichsten Studien dazu stammt von Vanheule, Desmet, Groenvynck, Rosseel und Fontaine (2008). Anhand eines Datensatzes von 404 ambulant behandelten Patienten und 695 Personen aus der Normalbevölkerung wurden zehn verschiedene in der Literatur beschriebene faktorenanalytische Modelle untersucht. Insgesamt wiesen alle Modelle ähnliche Kennwerte zur Passgüte auf, wobei vier Modelle – von Buckley et al. (2001), von Osman et al. (1997), von Viljoen et al. (2003) und von Ward (2006) – eine bessere Passgüte in beiden Samples (klinisch und nicht klinisch) aufwiesen als das Referenzmodell von Beck et al. (1996). Alle übrigen Modelle wiesen in mindestens einem Sample schlechtere Kennwerte als das Referenzmodell auf. Eine weitere Studie (Quilty, Zhang & Bagby, 2010) bestätigte gute Passgüten für das G-Faktor-Modell von Ward (2006) und das Dreifaktorenmodell von Osman et al. (1997) anhand eines Datensatzes ambulant behandelter Patienten mit einer major depressive disorder (MDD) nach DSM-IV.

Die vorliegende faktorenanalytisch orientierte Forschung kommt also nicht zu einem einheitlichen Resultat, führt sie doch zu zwei ganz unterschiedlichen Modellen an Symptomzusammenhängen in der Depression: Insbesondere das G-Faktor Modell und die obliquen, meist 2-faktoriellen Modelle (in der Folge als 2-Faktoren Modelle bezeichnet) unterscheiden sich inhaltlich erheblich. Es geht dabei nicht bloß um die Klärung, ob bei der häufig replizierten 2-Faktoren Struktur die affektiven Symptome dem kognitiven oder dem somatischen Faktor zuzurechnen seien. Vielmehr geht es um die auch praxisrelevante Frage, ob die Symptome der Depression in zwei gleichberechtigte Kategorien zerfallen oder ob im Wesentlichen *ein* depressives Syndrom existiert, dessen Symptome kognitive und somatische Komponenten aufweisen.

Während die Prüfung der psychometrischen Eigenschaften des BDI-II vor allem zu Beginn der Faktorenanalysen im Vordergrund stand (und selbstverständlich mit jeder Übersetzung wieder notwendig wird), rückte die Frage: „Ist die im BDI-II enthaltene Symptomatik und die Formulierung der Ausprägungsgrade adäquat, um das Konstrukt der Depression abzubilden?“ immer weiter aus dem Zentrum des Untersuchungsgegenstandes heraus. So hat die klinische Praxis diese Frage positiv beantwortet, denn inzwischen ist das BDI-II eines der meistverbreiteten Inventare zur Erhebung der Depression (Santor, Gregus & Welch, 2006).

Tabelle 1

Resultate der exploratorischen Faktorenanalysen des BDI-II (international)

Herleitung des Modells	Autoren	Faktor	Items
psychiatrische Stichprobe	Beck et al. (1996)	kognitiv	1, **2, 3, 5, 6, 7, 8, 9, 14**
		somatisch-affektiv	4, 10, 11, 12, 13, **15, 16,** 17, **18, 19, 20,** 21
psychiatrische Stichprobe	Steer et al. (1999)	kognitiv	**2, 3, 5, 6, 7, 8, 9, 14**
		somatisch-affektiv	1, 4, 10, 11, 12, 13, **15, 16,** 17, **18, 19, 20,** 21
psychiatrische Stichprobe	Arnau et al (2001)	kognitiv	**2, 3, 5, 6, 7, 8, 9,** 10, **14**
		somatisch-affektiv	1, 4, 8, 11, 12, 13, **15, 16,** 17, **18, 19, 20,** 21
psychiatrische Stichprobe	Bedi, Koopman & Thompson (2001)	kognitiv	**2, 3, 5, 6, 7, 8, 9, 14**
		somatisch-affektiv	1, 4, 10, 11, 12, 13, **15, 16,** 17, **18, 19, 20,** 21
psychiatrische Stichprobe	Buckley et al. (2001)	kognitiv	1, **2, 3, 5, 6, 7, 8, 9, 14**
		somatisch	11, **15, 16,** 17, **18, 19, 20,** 21
		affektiv	4, 10, 12, 13
psychiatrische Stichprobe	Osman, Kopper, Barrios, Gutierrez & Bagge (2004)	kognitiv-affektiv	1, **2, 3,** 4, **5, 6, 7, 8, 9,** 10, 12, 13, **14**
		somatisch	11, **15, 16,** 17, **18, 19, 20,** 21
psychiatrische Stichprobe	Viljoen, Grant, Griffiths & Woodward (2003)	kognitiv	**2, 3, 5, 6, 7, 8, 9,** 10, 13, **14**
		somatisch-affektiv	1, 4, 11, 12, 13, **15, 16,** 17, **18, 19, 20,** 21
psychiatrische Stichprobe	Keller, Hautzinger & Kühner (2008)	kognitiv	2, 3, 5, 6, 7, 8, 9, 14
		somatisch-affektiv	1, 4, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21
nicht-psychiatrische Stichprobe	Beck et al. (1996)	kognitiv-affektiv	1, **2, 3,** 4, **5, 6, 7, 8, 9,** 10, 11, 12, 13, **14,** 17, 21
		somatisch	**15, 16, 18, 19, 20**
nicht-psychiatrische Stichprobe	Osman et al. (1997)	negative Selbsteinstellung	1, **2, 3, 5, 6, 7, 8, 9,** 10, **14**
		Leistungsbeeinträchtigung	4, 12, 13, 15, 17, 19, 20
		somatisch	10, 11, **16, 18,** 21
nicht-psychiatrische Stichprobe	Dozois, Dobson & Ahnberg (1998)	kognitiv-affektiv	1, **2, 3, 5, 6, 7, 8, 9, 14**
		somatisch	4, 10, 11, 12, **15, 16,** 17, **18, 19, 20,** 21
nicht-psychiatrische Stichprobe	Keller et al. (2008)	kognitiv	1, 2, 3, 5, 6, 7, 8, 9, 14
		somatisch-affektiv	4, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21
theoretische Herleitung	Ward (2006)	Generell	Alle Items
		kognitiv	**2, 3, 5, 6, 7, 8, 9, 14**
		somatisch	**15, 16, 18, 19, 20**

Anmerkung. Fett gedruckte Items sind in allen Studien (mit Ausnahme der Studie von Osman et al., 1997) stichprobenunabhängig dem kognitiven bzw. dem somatischen Faktor zugeordnet.

Das unterstreicht seine potenzielle Nützlichkeit auch für die Theorie: Es ist entsprechend anzunehmen, dass das Konstrukt „Depression“ mit der Symptomliste des DSM-IV und der Messung durch das BDI-II hinreichend gut abgebildet wird (vgl. dazu auch Kühner, Bürger, Keller & Hautzinger, 2007).

Eine valide Abbildung der Depression ermöglicht nun, über die empirisch erhobenen Daten des Inventars auch die theoretischen Ansätze zum Störungsbild zu prüfen. Da die Analysen zur Struktur des BDI-II also auch eine Modellierung der Symptomstruktur des Konstrukts Depression selbst bedeuten, halten wir es für sinnvoll, die faktorenanalytisch nur unzureichend beantwortete Frage noch einmal aufzugreifen und mit einer anderen Analysemethodik anzugehen. Die Debatte weist schließlich auch starke praktische Implikationen auf, da immer wieder auf die unterschiedlichen Ansprechraten der verschiedenen Subgruppen auf spezifische Therapien hingewiesen wurde (z.B. Fink & Taylor, 2007; Stewart, Garfinkel , Nunes, Donovan & Klein, 1998). Eine der umfangreichsten Übersichten mit direkt davon abgeleiteten Therapieempfehlungen findet sich bei Damm, et al. (2009). Obwohl die nosologischen Betrachtungen in besagter Literatur über die reine Klassifikation nach symptomatischen Gesichtspunkten weit hinausgehen, konnte gezeigt werden, dass sich auch aus Unterschieden in der Symptomatik allein verschiedene Subgruppen der Depression ableiten lassen (Sullivan, Prescott & Kendler, 2002). Daher müsste man von einem Modell, das den Anspruch erhebt die Symptomstruktur des BDI-II (und also auch die Symptomstruktur der Depression) abzubilden, erwarten, dass dies in Übereinstimmung mit den breiteren Befunden zur Depressionsforschung steht. Unter Berücksichtigung der wachsenden Unzufriedenheit mit dem heterogenen Störungsbild der „Major Depression“, dem die Acta Psychiatrica Scandinavica 2007 eine volle Spezialausgabe gewidmet hat, müsste sich diese Heterogenität auch in den zugrundeliegenden Beeinflussungsfaktoren im BDI-II wiederfinden lassen.

Erstaunlicherweise weisen die faktorenanalytischen Modelle jedoch lediglich eine Unterscheidung der Symptomatik in kognitive und somatische (und z.T. affektive) Symptome auf, wobei die Frage bestehen bleibt, ob die Variabilität zwischen den kognitiven und somatischen Symptomen nun auf unterschiedliche Symptomausprägungen von Subgruppen depressiver Patienten oder auf residuale Kovarianzen der Kategorie (kognitiv/somatisch) zurückzuführen sind. Die starke Überlappung der Symptomatik bei allen Subgruppen ist auf alle Fälle für Faktorenanalysen (speziell exploratorische) schwer handhabbar. Auch wenn spezifische symptomatische Eigenheiten bei unterschiedlichen Subgruppen depressiver Patienten existieren, müssten sich diese substantiell voneinander unterscheiden, um in einem eigenen Faktor repräsentiert werden zu können, was offensichtlich in der vorhandenen Datenlage – zumindest beim BDI-II – nicht der Fall ist. Um diese feingliedrigen Symptomzusammenhänge zu entdecken, bieten sich andere strukturentdeckende statistische Methoden an. In der vorliegenden Studie schlagen wir deshalb zur Modellierung der Struktur des BDI-II (und damit zur Modellierung der Struktur der Depression) die Nonmetrische Multidimensionale Skalierung (NMDS) vor. Mithilfe der NMDS

kann die Struktur eines Inventars auf der Ebene der Items visualisiert werden, was im Detailgrad weit über die strukturentdeckenden Mittel der exploratorischen Faktorenanalyse hinausgeht.

In der vorliegenden Analyse werden die 21 Symptome des BDI-II anhand der Nonmetrischen Multidimensionalen Skalierung (NMDS) in einem Symptomraum abgebildet. In diesem Symptomraum entsprechen die (euklidischen) Distanzen zwischen den Symptomen der besten relationalen Abbildung der Ähnlichkeiten (gemessen als Pearson-Korrelationen) der Symptome untereinander. Liegen also Symptome im Symptomraum nahe beieinander, so treten diese gehäuft bei denselben Patienten und in ähnlicher Schweregradausprägung auf. Bei großer Distanz zwischen Symptomen treten sie (in Relation zu den anderen) selten bei denselben Patienten und/oder ähnlicher Schweregradausprägung auf. Da dies – im weiteren Sinne – auch das Ordnungskriterium bei exploratorischen Faktorenanalysen darstellt, lassen sich Faktorenstrukturen in einer NMDS-Lösung als Symptomcluster, also Symptome mit geringer Distanz zueinander, wiederfinden. Darüber hinaus folgt aber auch die Ordnung der Cluster untereinander den Ähnlichkeiten zwischen darin enthaltenen Symptomen. Sind sich also zwei Symptomcluster in Relation zu den anderen ähnlicher, so kommen diese ebenfalls näher zueinander zu liegen. Und es gibt noch eine Besonderheit in einer NMDS-Karte: Weist ein Symptomcluster die höchsten Ähnlichkeiten zu allen anderen Clustern auf, so wird dieses in der Mitte der Karte platziert, da an dieser Stelle die Distanzen zu den übrigen Clustern am geringsten sind.

Für die meisten Datensätze sind bereits 2-dimensionale Räume ausreichend, um die den Daten inhärente Struktur abzubilden, was eine „Symptomkarte“ ermöglicht und deren Interpretation begünstigt. Das Verfahren hat sich in der Darstellung von Symptomstrukturen psychopathologischer Inventare bereits als äußerst hilfreich erwiesen, wie Läge, Egli, Riedel und Möller (in Druck) anhand des AMDP-Inventars aufzeigen konnten.

Auch im Feld der Depressionsforschung ist die Multidimensionale Skalierung (von der die NMDS ein Spezialfall darstellt) nicht gänzlich neu. Cohen (2008) konnte anhand der Korrelationsmatrix für die Depressionsstichprobe aus Beck et al. (1996) mittels Multidimensionaler Skalierung im 2-dimensionalen Raum zeigen, dass sich eine 6-kategorielle Klassifikation der Depressionssymptome nach Beck, Rush, Shaw und Emery (1979) auch empirisch finden lässt und berichtete darüber hinaus Anhaltspunkte für eine – von Becks Klassifikation unabhängigen – „Arousal“-Dimension (mit 3 diskreten Ausprägungen), nach der sich die Symptome ordneten. Die MDS Lösung, also der entstandene 2-dimensionale Raum, wurde entsprechend anhand der resultierenden 3x6 Matrix („Arousal“ und Symptomkategorie) aufgeteilt und dahingehend interpretiert.

Des Weiteren haben Steinmeyer und Möller (1992) mittels NMDS-Analyse eine 2-dimensionale Facettenlösung der Hamilton-Depressionsskala zeigen können, die durch zwei

Ordnungsprinzipien – die Zentralität und die Lage auf einem bestimmten Kreissegment – charakterisiert werden konnte. Der Aspekt der Zentralität in der Lage der Symptome wurde als Indikator für den Schweregrad der Depression gesehen, die Lage auf unterschiedlichen Kreissegmenten für die Qualität der Symptomatik (Somatisation, Kognition, Verlangsamung und Schlaf). Analog hat Steinmeyer (1993) die klinische Validität der ersten Version des BDI mittels NMDS untersucht und gefunden, dass das BDI - besser als die HAM-D - intern und extern valide verschiedene klinisch bedeutsame Symptomkreise depressiver Erkrankungen erfasst, wenn auch die kognitive Seite übergewichtet ist und mehr auf Psychomotorik zielende Items fehlen.

Durch die grafische Darstellungsweise in einer 2-dimensionalen NMDS-Lösung lassen sich auch die faktorenanalytischen Modelle hervorragend vergleichen. Für eine stabile 2-faktorielle Struktur würde man zwei distinkte Cluster erwarten (die positive Grundkorrelation zwischen den obliquen Faktoren spielt in der relationalen NMDS-Lösung keine Rolle), während beim G-Faktor Modell drei Cluster – mit ihren Schwerpunkten auf einer Geraden angeordnet – erwartet würden. Die Pole sollten die spezifischen Symptome (kognitiv und somatisch), den Mittelpunkt die unspezifischen Symptome (vorwiegend affektiv) bilden. Dieser Gedanke wird im Methodik-Teil wieder aufgegriffen und im Detail beschrieben.

Das Ziel der vorliegenden Untersuchung ist entsprechend zweigeteilt. Erstens soll durch die Nonmetrische Multidimensionale Skalierung der Symptome im BDI-II eine Grundlage entstehen, auf der die divergierenden faktorenanalytischen Modelle miteinander verglichen werden können; und zweitens erhoffen wir uns durch die Abbildung der symptomatischen Struktur auf der Ebene der Symptome eine detailliertere Modellierung der Depression als dies bislang in den faktorenanalytischen Modellen der Fall ist.

Methodik

Stichproben

Als Datengrundlage für die vorliegende Untersuchung diente die Stichprobe, welche auch dem Manual der deutschsprachigen Version des BDI-II zugrunde liegt (Hautzinger et al., 2006). Verwendet wurden ausschließlich depressive Patienten (N = 266), welche im Rahmen von stationären und ambulanten Routinebehandlungen in unterschiedlichen Kliniken und Therapiezentren Deutschlands erhoben wurden. Das Durchschnittsalter der Stichprobe lag bei 48.8 Jahren (SD = 15.7), der Frauenanteil betrug 65.4%.

Statistische Methoden

Zur Analyse der Daten wurde die Nonmetrische Multidimensionale Skalierung (NMDS) des Softwarepakets ProDaX (Oberholzer, Egloff, Ryf, & Läge, 2008) verwendet. In der NMDS werden Objekte auf der Grundlage von Proximitäten (jedes Objekt von *n* Objekten besitzt *n-1* Proximitäten zu den übrigen Objekten) in einen euklidischen Raum abgebildet. Die Berechnung der optimalen Konfiguration wird durch einen iterativen Algorithmus vorgenommen, der die Proximitätsmatrix in Distanzrelationen der Objekte umsetzt und diese möglichst rangtreu (bei der NMDS, bei der MDS intervalltreu) in einen niedrig-dimensionalen Raum abbildet. Da eine solche Abbildung bei realen Daten und niedrig dimensionalen Räumen praktisch nie perfekt möglich ist, müssen Abweichungen zwischen einer idealen (strikt rangtreuen) und einer realen (möglichst rangtreuen) NMDS-Lösung in Kauf genommen werden. Als Maß für die Passgüte der Abbildung steht der Stresswert (Borg & Groenen, 2005). In der vorliegenden Analyse wurde ein robuster Berechnungsalgorithums (Robuscal) eingesetzt (Läge, Daub, Bosia, Jäger & Ryf, 2005), um den Einfluss verrauschter Daten so gering wie möglich zu halten.

Als Proximitätsmaß diente die Pearson-Korrelation. Die Grundlage zur Berechnung des Symptomraumes stellte entsprechend die Korrelationsmatrix der Symptome, d.h. die paarweisen Korrelationskoeffizienten der Symptome des BDI-II dar. Durch die Verwendung der Korrelationsmatrix zur Berechnung des Symptomraums lassen sich in der Regel faktorenanalytisch gefundene Strukturen in der NMDS-Lösung als Cluster von Symptomen wiederfinden (und zwar unabhängig von der Dimensionalität der NMDS-Lösung). Der folgende Gedankengang zeigt, weshalb dies so gefunden wird:

Es sei ein Datensatz gegeben, dessen wahre Struktur faktorenanalytisch gut erfasst werden kann (kategoriale Struktur) und dessen Faktoren – vorerst – unabhängig voneinander seien. In diesem Datensatz existieren also verschiedene Gruppen von Items, welche innerhalb der Gruppen untereinander hochgradig korreliert, zwischen den Gruppen, entsprechend der geforderten Unabhängigkeit der Faktoren, bis auf zufällige Zusammenhänge unkorreliert sind. Eine exploratorische Faktorenanalyse wird in diesem Datensatz von *n* Items mit *k* Gruppen von Items *k* Faktoren finden, wobei jeder Faktor die Symptome der *k* Gruppen bestmöglich – das heißt mit maximaler Kovarianz zu allen Symptomen innerhalb der Gruppe – repräsentiert. Die NMDS-Lösung wird entsprechend aus *k* Clustern bestehen, welche möglichst äquidistant im Raum verteilt zu liegen kommen werden. Cluster werden sich zeigen, da die Korrelationen zu gruppeneigenen Items höher sind als zu gruppenfremden (was eine höhere Ähnlichkeit und entsprechend eine geringere Distanz zur Folge hat). Äquidistanz im Raum wird sich einstellen, da die Korrelationen zu allen gruppenfremden Items dieselben sind, bei postulierter Unabhängigkeit also null. Eine streng kategoriale Lösung, in der mit wenigen Faktoren ein Großteil an Varianz in den Daten erklärt werden kann, wird sich in der NMDS-Lösung demzufolge als klar distinkte Item-Cluster zeigen mit nur geringen Intra-Cluster-Distanzen. Handelt es sich um eine Lösung

in der eher heterogene Konstrukte durch Gruppen von Variablen erklärt werden, so sind entsprechend größere Intra-Cluster-Distanzen zu erwarten.

Durch die relationale Darstellung der Items zueinander bleibt allerdings ein nicht unwesentlicher Aspekt in den Daten unberücksichtigt: Eine Korrelationsmatrix mit vorwiegend positiven Korrelationen – d.h. wenn die Symptome allesamt positiv miteinander korreliert sind – wird in der NMDS-Lösung genau gleich repräsentiert, wie wenn die Korrelationsmatrix vorgängig am Mittelwert zentriert würde. Dieser Umstand muss zwingend bei der Interpretation von NMDS-Lösungen berücksichtigt werden. Die relationale Behandlung der Ähnlichkeiten zwischen den Items macht auch die vorab geforderte Unabhängigkeit zwischen den Faktoren überflüssig. Es werden nur diejenigen Varianzen berücksichtigt, die auch auf unterschiedliche Beeinflussungen der Faktoren zurückgehen. Positive Korrelationen zwischen den Faktoren, wie dies in den obliquen Faktorenmodellen zugelassen wird, zeigen sich in der NMDS-Lösung nicht.

Die obengenannten Ordnungskriterien für NMDS-Lösungen treffen gezielte Vorhersagen, wie sich die Symptomstruktur auf der Ebene der Symptome zeigen müsste, bei adäquater Modellierung der BDI-II Daten durch die unterschiedlichen faktorenanalytischen Modelle. Eine 2-Faktoren Struktur, die für alle Symptome ausschließlich einen beeinflussenden Faktor annimmt (z.B. für die kognitiven Symptome den kognitiven Faktor, für die somatisch-affektiven Symptome den somatisch-affektiven Faktor), müsste sich in der NMDS-Lösung als klar separierte 2-Cluster Lösung zeigen. Die Kovarianzen der kognitiven Symptome wären ausschließlich auf die Beeinflussung des kognitiven Faktors zurückzuführen, diejenigen der somatisch-affektiven Symptome auf die Beeinflussung des somatisch-affektiven Faktors. Die beiden Cluster müssten voneinander in großer Distanz (im Vergleich zu den Distanzen innerhalb des Clusters) zu liegen kommen, da kein Symptom von beiden Faktoren beeinflusst wird. Die G-Faktor Struktur würde sich grob in drei Symptomcluster gliedern (kognitiver-, somatischer- und G-Faktor). Der G-Faktor spielt in der relationalen NMDS-Lösung deshalb weiterhin eine Rolle, weil einige Symptome exklusiv vom G-Faktor beeinflusst werden und damit eine Art Zentrum bilden. Im G-Faktor Modell werden einzelne Items von mehreren Faktoren beeinflusst – die kognitiven Items vom G-Faktor und dem kognitiven Faktor und die somatischen Items vom G-Faktor und dem somatischen Faktor. Diese Symptome können entsprechend als Linearkombination der beiden Faktoren aufgefasst werden, wobei jedem Symptom unterschiedliche Gewichte für die beeinflussenden Faktoren zugewiesen werden können. Damit lässt das G-Faktoren Modell eine dimensionale Ordnungskomponente zu. Die drei Symptomgruppen müssten sich also auf einer Geraden anordnen, mit denjenigen Symptomen in der Mitte, die ausschließlich vom G-Faktor beeinflusst werden. Durch die dimensionale Ordnungskomponente besitzt dieses Modell die Freiheit, die Distanzen zwischen den 3 Clustern auf ein Minimum zu reduzieren, wobei im Extremfall ein fließender Verlauf in drei (allerdings immer noch strikt voneinander getrennte) Bereiche entstehen kann.

Da für jedes Faktorenmodell spezifische Vorhersagen zur Strukturierung der Symptome in der NMDS-Lösung getroffen werden können, lassen sich auf deren Grundlage die unterschiedlichen faktorenanalytischen Modelle bereits in einer 2-dimensionalen Darstellung der NMDS miteinander vergleichen.

Ergebnisse

Abbildung 1 zeigt die NMDS-Lösung der BDI-II Symptome auf der Datengrundlage der 266 depressiven Patienten. Der Stresswert von 0.22 zeigt eine – für die Anzahl an abgebildeten Objekten – akzeptable Einpassung in den 2-dimensionalen Raum an (Gigerenzer, 1981). Zur Semantik fällt auf, dass die somatischen Items „Weinen“, „Unruhe“ und „Reizbarkeit“, sowie „Verlust an sexuellem Interesse“, „Appetitveränderung“ und „Schlaf“ zwei voneinander und von den übrigen Items distinkte Cluster bilden. Die kognitiven Symptome „Suizidgedanken“ und „Bestrafungsgefühle“ positionieren sich ebenfalls etwas abseits. Eine weitere Auffälligkeit stellen die affektiven Items „Traurigkeit“, „Entschlusslosigkeit“, „Verlust an Freude“ und „Interessenverlust“ und das Symptom „Energieverlust“ in der Mitte der Karte dar. Deren zentrale Lage deutet auf eine hohe Ähnlichkeit mit allen übrigen Symptomen hin. Insgesamt lässt die Karte eine klare Clusterung (insbesondere eine klare Zweiteilung) der Symptome aber vermissen – eine kategoriale Lösung drängt sich nicht direkt auf.

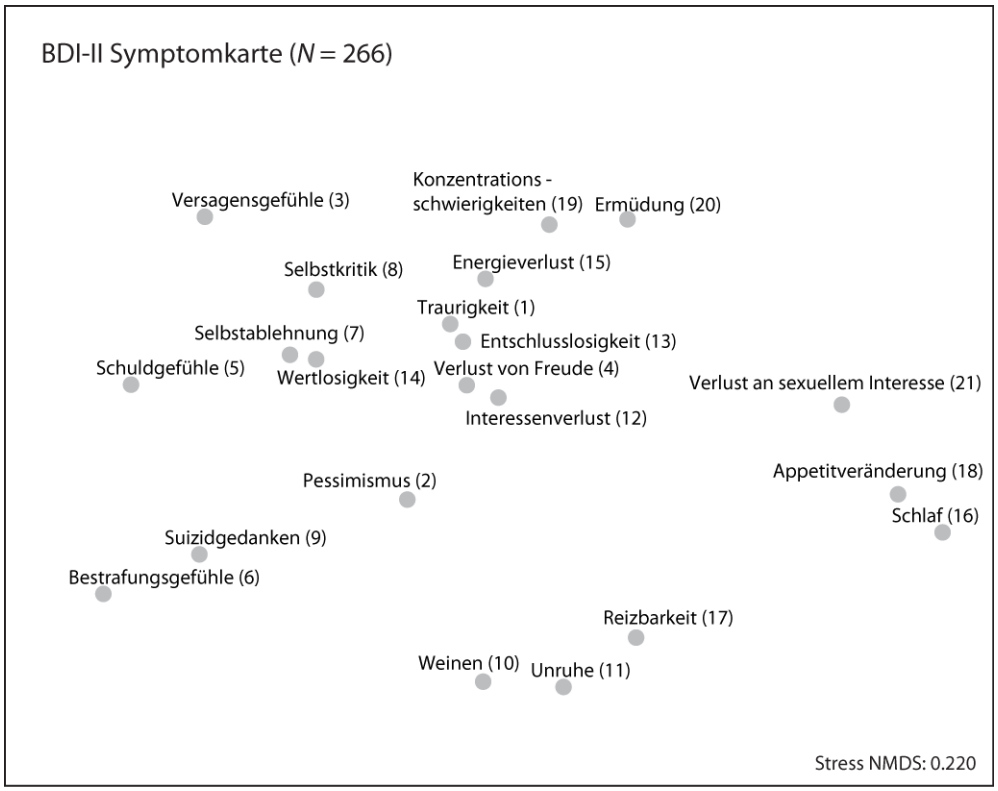

Abbildung 1. NMDS-Lösung der Symptomstruktur der BDI-II Items in einer Stichprobe von depressiven Patienten (N=266).

Bei genauerer Betrachtung der Items in Relation zu ihren Positionen auf imaginierten x- und y-Achsen stellt man fest, dass sich die kognitiven Symptome im linken Bereich der Karte, die somatischen vorwiegend im rechten Bereich der Karte und die affektiven Symptome etwa in der Mitte der Karte positionieren. Wenn diese auch nicht klar als Gruppen voneinander abgrenzbar sind, so lassen sie sich doch als Regionen in der Karte wiederfinden. Die 2-Faktoren Modelle klären vorwiegend die Varianz entlang der x-Achse auf – was sich auch in der Namensgebung der Faktoren widerspiegelt. Da die Distanz zwischen den äußersten Symptomen auf der x-Achse größer ist als diejenige der äußersten Symptome auf der y-Achse (was einem elliptischen, entlang der x-Achse gestreckten Punkteschwarm entspricht), ist entlang der x- Achse auch mehr Varianz aufzuklären als entlang der y-Achse. Trotzdem zeigt sich entlang der y-Achse in Abbildung 1 noch ein substantieller Anteil an Varianz mit systematischem Informationsanteil. Eine dimensionale Struktur ist hier zwar nicht durchgängig vorhanden. Im oberen Bereich der Karte positionieren sich überwiegend Symptome die mit einem geringen Aktivitätsniveau einhergehen, während im unteren Bereich der Karte eher Symptome zu finden sind, welche mit einem hohen Aktivitätsniveau assoziiert sind. Während diese Interpretation im mittleren Bereich des x-Achsenabschnitts gut passt, stimmt sie in den Randbereichen nur begrenzt. Im Diskussionsteil stellen wir deshalb eine Facettenlösung vor, die uns für die Interpretation der NMDS-Lösung adäquater erscheint.

Abbildung 2 zeigt zusätzlich zur NMDS-Lösung vertikal bzw. horizontal schraffiert die kategoriale Zuordnung der Items zum kognitiven und zum somatisch-affektiven Faktor, so wie diese im vorliegenden Datensatz mittels Promax-Rotation und zweifaktorieller Lösung gefunden werden und in Keller et al. (2008) ausführlich beschrieben sind. Die Abbildung der Symptome in der Karte folgt der Befundlage aus Tabelle 1 zum kognitiven und zum somatischen Faktor: Symptome, welche einheitlich dem kognitiven Faktor zugeordnet wurden sind als Quadrate, Symptome, welche einheitlich dem somatischen Faktor zugeordnet wurden als Dreiecke, und Symptome, welche uneinheitlich zugeordnet wurden sind als Kreise dargestellt.

In der Symptomkarte von Abbildung 2 können der kognitive und der somatisch-affektive Faktor gemäß der Zuordnung bei Keller et al. (2008) anhand der x-Achse voneinander abgegrenzt werden. Allerdings wird auch deutlich, dass insbesondere die affektiven Items „Traurigkeit“, „Entschlusslosigkeit“, „Verlust an Freude“ und „Interessenverlust“, als auch die Items „Pessimismus“, „Energieverlust“ und „Weinen“, welche in etwa auf halbem Weg der x-Achse zu liegen kommen, sich nur in knapper Distanz zur Trennlinie zwischen den beiden Faktoren positionieren, was eine Zuordnung zum einen oder anderen Faktor aus Sicht der NMDS-Lösung unsicher macht.

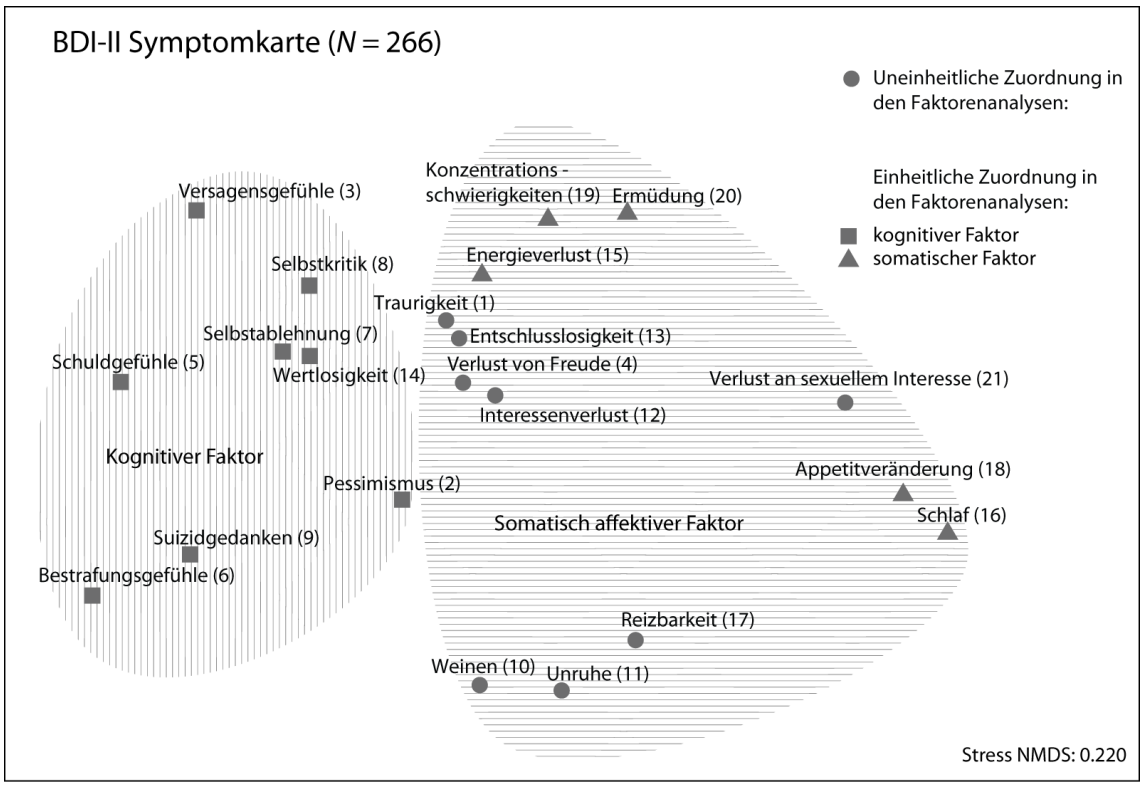

Abbildung 2. Symptomstruktur des BDI-II in einer Stichprobe von depressiven Patienten (N=266). vertikal bzw. horizontal schraffiert sind die Faktoren gemäß exploratorischer Faktorenanalyse des zugrundeliegenden Datensatzes. Die Quadrate bzw. Dreiecke zeigen den größten gemeinsamen Nenner der exploratorischen Faktorenanalysen in der Literatur (vgl. dazu Tabelle 1).

Für die in der Literatur einheitlich zugeordneten Symptome zeigt sich der kognitive Faktor als robust (mit 8 beständig zugeordneten Symptomen) und kann auch in der NMDS-Lösung als Facette identifiziert werden – allerdings mit beträchtlicher Varianz zwischen den zusammengefassten Symptomen. Der somatische Faktor dagegen, der unter Ausschluss der Studie von Osman et al. (1997) noch fünf Symptome umfasst, kann in der NMDS-Lösung nicht als Facette identifiziert werden. Dafür ist die Lage der fünf Symptome zu unterschiedlich, und das nicht enthaltene Symptom „Verlust an sexuellem Interesse“ befindet sich zwischen den beiden Gruppen von Symptomen „Appetitveränderung“ / „Schlaf“ und „Energieverlust“ / „Konzentrationsschwierigkeiten“ / „Ermüdung“.

Diskussion

Die NMDS-Lösung der BDI-II Symptome (Abbildung 1) zeigt nur ansatzweise eine kategoriale Lösung. Die kognitiven und affektiven Symptomgruppen können zwar strikt voneinander getrennt werden, der Übergang zwischen den Faktoren zeigt sich in der NMDS-Lösung allerdings als unscharf (geringe Distanz zwischen den „Rand-Symptomen“ der jeweiligen Faktoren). In

einem obliquen 2-Faktoren Modell ist die Zuordnung der Symptome nahe der Trennlinie (Abbildung 2) zu einem der beiden Faktoren entsprechend unsicher. Betroffen sind vorwiegend die affektiven Symptome, was die stichprobenabhängigen Resultate bei deren Zuordnung in den 2-Faktoren Modellen erklärt.

Gerade bei den mittig positionierten Symptomen, welche in ihrer Zuordnung stichprobenabhängig reagieren, handelt es sich allerdings um Kernsymptome der Depression (Damm, et al., 2009). Der Fokus der beiden Faktoren (kognitiv und somatisch) liegt aber eher in den Randbereichen der NMDS-Lösung und damit bei den spezifischeren Symptomen der Depression. Nun stellt sich die Frage, inwieweit eine Zweikomponenten-Kategorisierung sachdienlich ist, wenn der zentrale Aspekt quasi nur „mitkategorisiert“ wird. Die Bildung zweier Subskalen von BDI-II Items auf der Grundlage von obliquen zwei Faktorenlösungen ist vor dem Hintergrund der NMDS-Lösung entsprechend abzulehnen.

Die durchwegs positive mittlere Korrelation zwischen den Symptomen ($\bar{r} = 0.4, \quad \hat{\sigma}_r = 0.12$) und die gesamthaft wenig kategoriale NMDS-Lösung deutet eher in Richtung eines unspezifischen depressiven Syndroms, wie dies vom G-Faktor Modell konstatiert wird (Ward, 2006). Der Einbezug von spezifischen kognitiven und somatischen Symptomen und deren Zuordnung zu den jeweiligen Faktoren im Modell hilft, die verbleibenden (orthogonalen) maximalen Restvarianzanteile zu bestimmen und zuzuordnen. Vor dem Hintergrund der NMDS-Lösung stellt das G-Faktor Modell von Ward (2006) also das am besten passende Faktorenmodell dar, obschon es einiges an Variabilität (speziell entlang der y-Achse) unerklärt lässt. Offen bleibt zudem die Frage, ob diese maximalen Restvarianzanteile aufgrund von unterschiedlichen Subtypen der Depression (Personengruppen) oder aufgrund bestehender Kovarianzen zwischen Symptomen derselben Kategorie (kognitiv/somatisch) entstehen. Die Beantwortung dieser Frage geht über die hier vorliegende Untersuchung hinaus, könnte aber vielleicht über eine Latent Class Analyse untersucht werden - ähnlich der Untersuchung von Sullivan et al. (2002), allerdings beschränkt auf eine Stichprobe von depressiven Patienten.

Die NMDS-Lösung zeigt zusätzlich zur Variabilität zwischen kognitiven und somatischen Symptomen erhebliche (und in den Faktorenanalysen unerklärte) Varianz entlang der y-Achse. Diese Strukturanteile werden in den Faktorenanalysen selten berücksichtigt, da sie alleine nicht mehr genügend Varianz für einen eigenen Faktor enthalten. (In einer Hauptkomponentenanalyse des vorliegenden Datensatzes beispielsweise musste eine dritte Hauptkomponente als nicht mehr interpretierbar zurückgewiesen werden, vgl. Keller et al., 2008). In der Multidimensionalen Skalierung dagegen scheint das Ordnungskriterium eines Aktivitätsniveaus bzw. eines „Arousals“, zwar nicht auf Item-, aber doch auf Konzeptebene stabil: Sowohl in der hier vorgelegten Arbeit als auch in der Arbeit von Cohen (2008) konnte ein solches entlang der y-Achse identifiziert werden. Auch bezüglich der Anordnung der Symptome entlang der x-Achse sind die beiden Lösungen – bis auf die Symptome Energieverlust und Suizidgedanken –

in identische Regionen unterteilbar. Zur Interpretation der MDS-Lösung zieht Cohen (2008) eine Einteilung des Raums in 18 Regionen (3 „Arousal“ mal 6 Symptomkategorien) heran, vor dessen Hintergrund die Platzierung der Symptome interpretiert wird. Damit verliert er allerdings die durch Multidimensionale Skalierung gewonnene dimensionale Sicht auf die Lösung und bewegt sich wieder auf einer (freilich detaillierteren) rein kategorialen Interpretationsebene.

Steinmeyer und Möller (1992) dagegen legen in ihrer NMDS-Lösung des HAMD gerade auf die Dimensionalität, also auf die Interpretation einer Region im Kontext der Gesamtstruktur, grosses Gewicht. Die Kernsymptome der Depression im HAMD fanden sie in ihrer Lösung mittig positioniert und vermuteten diese als Symptome mit maximaler Schweregradabhängigkeit. Äquivalent dazu die vorliegende NMDS-Lösung des BDI-II: Auch hier zeigen sich die Kernsymptome der Depression in der Mitte der Struktur (Abbildung 3).

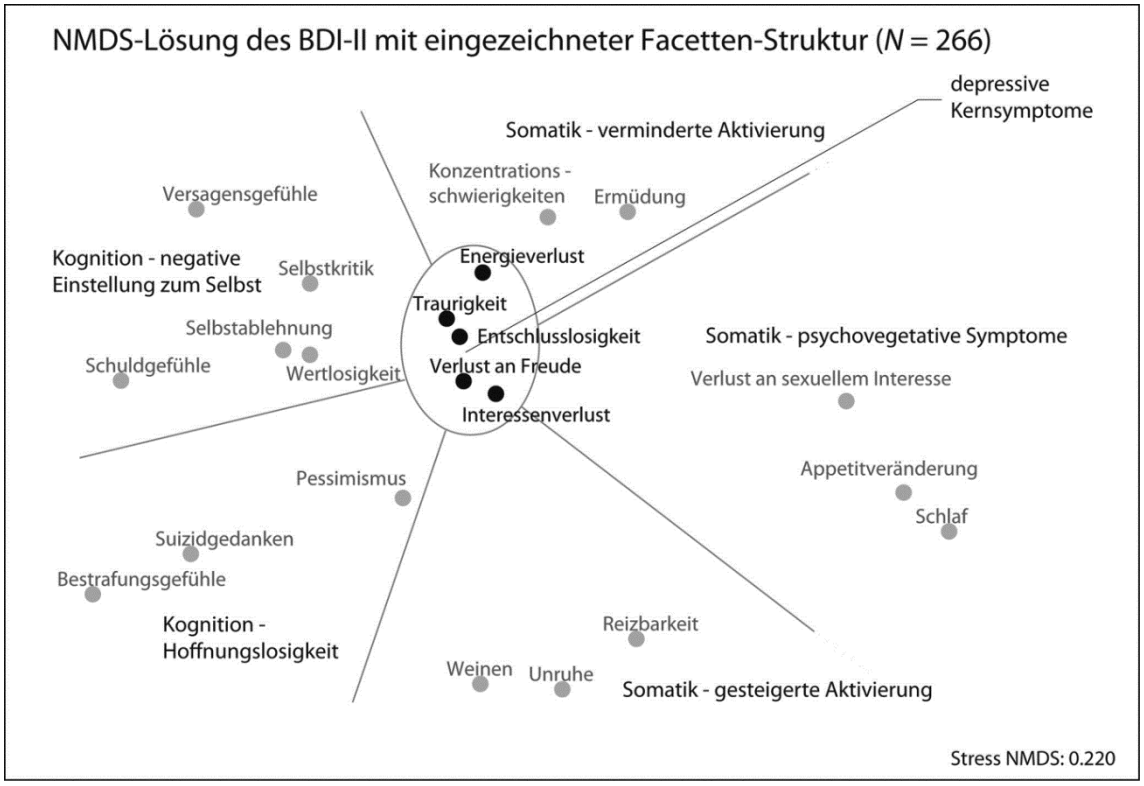

Abbildung 3. Facettenlösung des BDI-II. Während die affektiven Symptome in Richtung Mitte platziert werden, differenzieren gegen außen hin die Facetten „verminderte Aktivierung“, „psychovegetative Symptome“, „gesteigerte Aktivierung“, „Hoffnungslosigkeit“ und „negative Einstellung zum Selbst“ die Symptomatik unterschiedlicher Störungsbereiche.

Rund um diese Kernsymptome lässt sich eine Facettenstruktur finden, die selbst auch wieder eine dimensionale Ordnung aufweist: Von 12 Uhr im Uhrzeigersinn beginnend, über die vorwiegend somatischen Facetten „verminderte Aktivierung“, „psychovegetative Symptome“ und „gesteigerte Aktivierung“ (wobei „gesteigerte Aktivierung“ und „verminderte Aktivierung“ die Antipoden der y-Achse bilden), zu den kognitiven Facetten der „Hoffnungslosigkeit“ und der „negativen Einstellung zum Selbst“ (wobei die Gruppen somatische und kognitive Symptome die Antipoden der x-Achse bilden). Die Facetten kategorisieren die Symptome entsprechend ihren zentralen Merkmalen innerhalb der Gesamtstruktur. Das von Cohen (2008) als dimensionales Ordnungskriterium vorgeschlagene „Arousal“ finden wir aber stärker kategorial geprägt. Im mittleren x-Achsenabschnitt herrscht zwar noch mehrheitlich eine dimensionale Ordnung der Symptome, gegen die linken und rechten Aussenbereiche der Karte hin wird diese Ordnung aber sichtlich schwächer. Statt einer „Arousal“-Dimension, die ein globales, für den gesamten Symptomraum gültiges Ordnungskriterium darstellen würde, schlagen wir deshalb ein lokaleres „Arousal“-Ordnungskriterium auf der Basis der beiden Aktivierungsfacetten vor. Die äußeren Facetten grenzen die spezifischeren Symptome – welche als Diskriminierungsgrundlage bei unterschiedlichen Subtypen der Depression herangezogen werden könnten – voneinander und gegen die Kernsymptome hin ab. In der Mitte der NMDS-Lösung vermuten wir generelle, der Depression inhärente Symptome, die bei allen Subtypen der Depression gefunden werden und eine maximale Schweregradabhängigkeit aufweisen.

Die Facetten zeigen, dass sich innerhalb der Symptomatik der Depression ausgeprägte, semantisch interpretierbare Varianzquellen zeigen. Dies begünstigt die Interpretation, dass sich innerhalb des Störungsbilds der Depression nach symptomatischen Gesichtspunkten unterscheidbare Subtypen bilden lassen. Die Facettenlösung des BDI-II unterstützt damit die Forderung nach einer feingliedrigeren Unterteilung der Depression (z.B. Damm, et al., 2009; Fink & Taylor, 2007; Joiner et al., 2005; Parker, 2007; Stewart et al., 2007; Shorter, 2007). Die bislang von den Faktorenmodellen ignorierte Komponente der Aktivierung zeigt sich als klar ausgeprägtes Ordnungskriterium zwischen den Facetten der „verminderten-„ und der „gesteigerten Aktivierung“. Diese systematische Varianzquelle in den Daten, die immer wieder mit der Klassifikation von Subtypen in Verbindung gebracht wird, zeigt sich ausschließlich in der sensitiveren NMDS-Lösung.

Die reine Zerlegung der Depressionssymptome in eine kognitive und somatische Komponente scheint vor dem Hintergrund der NMDS-Lösung entsprechend als aufgezwungen. Obwohl damit die hauptsächliche varianzstiftende Quelle – nach dem generellen Faktor – erfasst werden kann, wird es der Struktur der Daten nicht gerecht. Die Ordnung der Symptome ist eher dimensionaler als kategorialer Natur. Damit deckt eine Facettenlösung die praktischen und theoretischen Anforderungen an eine Kategorisierung wesentlich besser ab und wird von den Daten klar unterstützt. Die geringe Anzahl an beteiligten Symptomen pro Facette lässt die Bildung von

Subskalen allerdings auch auf der Grundlage der Facettenlösung als wenig zweckdienlich erscheinen (zumindest wenn es um die Bestimmung des Schweregrads einer Erkrankung geht). Das führt dazu, dass der an sich hohe Detailgrad an Information, der in den Daten des BDI-II grundsätzlich erfasst wird, leider noch nicht vollumfänglich erschlossen werden kann.

Einschränkend ist anzumerken, dass es sich bei der hier vorgestellten Facettenlösung des BDI-II um eine Auswertung exploratorischen Charakters handelt. Eine Validierung anhand einer unabhängigen Stichprobe wäre wünschenswert – insbesondere im Hinblick auf die Stabilität der Facettenlösung. Mit der MDS-Lösung von Cohen (2008) liegt zwar potentiell eine Struktur vor, auf die hin geprüft werden könnte, in drei Belangen unterscheiden sich aber die beiden Studien: Erstens wurden zwei unterschiedliche Algorithmen (im Speziellen zwei verschiedene Minimierungsfunktionen) verwendet, zweitens handelt es sich bei Cohen (2008) um eine gemischte Stichprobe von ambulant behandelten Depressiven und Patienten mit anderen affektiven Störungen (knapp die Hälfte der Stichprobe von Beck et al. (1996) weisen die Diagnose einer Angst-, Anpassungs- oder anderen Störung auf), während in der vorliegenden Studie nur depressive Patienten untersucht wurden und drittens stammen die Daten aus der englischen Version des BDI-II. Die Attribuierung der Unterschiede in den beiden Lösungen auf spezifische Faktoren wird entsprechend unsicher. Weiter ist anzumerken, dass mit dem BDI-II ein reines Selbsterfassungsinventar vorliegt; ob und wie weit sich die hier gefundene Facettenstruktur auch in einer Aussensicht auf die Depression replizieren lässt, geht über den Rahmen dieser Studie hinaus.

Immerhin wird durch die NMDS die symptomatische Struktur des BDI-II erstmalig derart abbildbar, dass auch wenig erfahrene Diagnostiker die komplexen Symptomverflechtungen erkennen können. Eine Darstellung des BDI-II Befundes direkt in einer NMDS-Lösung (also eine Visualisierung des individuellen Symptomprofils durch schweregradabhängige Einfärbung der Symptome) könnte damit auch praktische Relevanz aufweisen: Die farbliche Kodierung des Schweregrades der Einzelsymptome liesse auf einen Blick allenfalls vorherrschende Problembereiche erkennen – nämlich als Anhäufungen von hohen Symptomschweregraden innerhalb spezifischer Facetten. Die Interpretationsgrundlage bliebe dabei aber die individuell eingefärbte Symptomkarte, womit auch die dimensionale Ordnung der Symptome nicht aus den Augen verloren werden kann. Die bislang primäre Betrachtung des numerischen Schweregrades würde also um eine bildliche Darstellung des Symptomprofils ergänzt, dessen Analyseebene sich nicht weiter auf aggregierte Werte, sondern direkt auf die Symptome respektive die Facetten bezöge. Im Grundsatz fände sich so eine Darstellungsform, welche sowohl quantitative als auch qualitative Aspekte der Depression berücksichtigen würde und damit – als praktische Anwendung – die Diagnostik und eine darauf aufbauende Behandlungsplanung möglicherweise vereinfachen könnte.

Activation as an overlooked factor in the BDI-II: a factor model based on core symptoms and qualitative aspects of depression

Joël Bühler[1], Ferdinand Keller[2], Damian Läge[1]

[1]Department of Psychology, University of Zurich, Switzerland
[2]Department of Child and Adolescent Psychiatry and Psychotherapy, University of Ulm, Germany

Submission status:

Resubmitted to *Psychological Assessment,* July 2013.

Authors' contributions:

Joël Bühler: Development of the research question, review of the literature, execution of NMDS analyses, interpretation of the results, writing of the manuscript

Ferdinand Keller: Provision of the data, execution of factor analyses, revision of the manuscript

Damian Läge: Supervision and discussion of Joël Bühler's contributions, revision of the manuscript

Abstract

An adequate assessment of depression has been of concern to many researchers over the last half-century. These efforts have brought forth a manifold of depression rating scales, of which the Beck Depression Inventory (BDI) is one of the most commonly used self-assessment scales. Since its revision, the item structure of the BDI-II has been examined in many factor-analytic studies, yet a consensus about the underlying factor structure could not be achieved. Recent findings from a Nonmetric Multidimensional Scaling (NMDS) analysis (Bühler, Keller, & Läge, 2012) of the German norming sample of the BDI-II emphasized a structure with different qualitative aspects of depression, which suggested that the existing factor models do not adequately represent the data. The NMDS results were reviewed and, based on these findings, a different factor model is proposed. In contrast to the common factor models in the literature, the presented model includes an additional factor, which is associated with the activation level of the BDI-II symptoms. The model was evaluated with a second sample of patients diagnosed with a primary affective disorder (N = 569) and obtained good fit indices that even exceeded the fit of the most reliable factor model (Ward, 2006) described in the literature so far. Additionally, emphasis is laid on the methodological question, how factor models may be derived from the results of NMDS analyses.

Keywords: Beck Depression Inventory-II, confirmatory factor analysis, Nonmetric Multidimensional Scaling, depressive patients

Introduction

The Beck Depression Inventory (BDI; Beck, Ward, Mendelsohn, Mock, & Erbaugh, 1961) is a wide spread self-assessment questionnaire to measure the severity of depression. It consists of 21 items that assess a wide range of depressive symptoms. Each item has four categories that are formulated in a symptom-specific way; the total score of these items reflect the severity of depression. In 1996, a minor revision of the BDI was carried out to meet the criteria of the DSM-IV (American Psychiatric Association, 1994) and resulted in the BDI-II (Beck, Steer, & Brown, 1996). The revision of the BDI-II has given rise to repeated psychometric evaluation to ensure the quality of the test. Extensive research has been conducted on its item structure, mainly by applying factor analytic techniques. However, previous results on the factor structure of the BDI-II are diverging.

Structure of the article and aims of the current study

To facilitate readability and understanding of the following introductory paragraphs, we chose to present a short overview on the structure of the text and on the aims of the study at the beginning of the article. The current article pursued two equally pronounced goals: Firstly, a new factor model of the BDI-II is proposed. This newly proposed model includes an additional factor, which is argued to relate to the level of activation of the BDI-II symptoms. The model was derived from a Nonmetric Multidimensional Scaling (NMDS) solution of the BDI-II (Bühler et al., 2012) and was evaluated with an independent sample of depressive patients. Secondly, the general procedure of deriving factor models from NMDS solutions is demonstrated. This procedure can be applied to virtually any questionnaire and thus is not limited to the BDI-II or the domain of psychopathological inventories.

The introduction is structured to consider both goals in separate paragraphs. In the first paragraph, a dichotomization of factor models is described which separates two structurally different groups of models in the literature (simple and complex factor models). The dichotomization is especially useful to separate the models' representations in NMDS solutions. In a second paragraph, previous factor analytic findings concerning the item structure of the BDI-II are reviewed. In a third paragraph, previous NMDS findings of the BDI-II and the endorsement of an activation factor in depression and in the BDI-II items are presented. In a fourth paragraph, because NMDS is a rarely used analysis method, the key concepts necessary to understand the representation of factor models in NMDS solutions are explained. Lastly, an NMDS solution of the BDI-II (Bühler et al., 2012) is examined with respect to the plausibility of the two different types of factor solutions (simple and complex). Furthermore, the process to derive the proposed factor model is explained. Readers solely interested in the results of the model proposed may skip the fourth and the fifth subsection of the introduction and may directly continue with the methods section.

A dichotomy of factor models

The factor analyses in previous studies have been conducted on two different methodological premises, which resulted in structurally different factor models of the BDI-II. The dichotomy to categorize previous factor models that was applied in this article is based on the terminology originally coined by Thurstone (1954) and distinguishes between simple and complex factor structures.

Simple factor structures are indicated, if each item is associated with at most one factor. Simple factor structures are usually obtained in exploratory factor analysis due to rotation criteria that favor simple factor structures to solve the indeterminacy problem in factor analyses. The models implementing simple factor structures of the BDI-II are referred to as simple factor models in the remainder of this article. Simple factor models comprise models applying different rotation criteria, different number of factors and different constraints on the orthogonality of the postulated factors. The models commonality is found in the composition of the items' variances: the variance of each item comprises the variance explained by exactly one associated factor (explained variance) as well as error variance (unexplained variance). Thus, simple factor models are essentially dichotomic: either an item is associated with a factor or it is not.

In contrast, complex factor structures are indicated, if some items are associated with at least two different factors. A special case of complex factor structures are bi-factor structures, in which each item is associated with at most two factors – the general factor and one of the group factors (e.g. Jennrich & Bentler, 2011). The models implementing a complex factor structure of the BDI-II are referred to as complex factor models in the remainder of this article. Complex factor models comprise models applying different number of factors, different item-factor association patterns and different constraints on the orthogonality of the postulated factors. Complex factor models share the feature that they are able to explain an item's variance by multiple components. Depending on the number of factors associated with an item, the item's variance can be explained as a linear combination of the associated factors, allowing for interpretable differences in the item's factor loading patterns.

Recent factor analytic findings of the BDI-II

The majority of studies proposed a simple factor model of the BDI-II. The findings of these studies are highly heterogeneous though, as regards the associations of items to factors. However, two association patterns have repeatedly been found: either a cognitive and a somatic-affective, or a cognitive-affective and a somatic factor were obtained. These patterns have been replicated in most of the following studies (e.g. Arnau, Meagher, Norris, & Bramson, 2001; Keller, Hautzinger, & Kühner, 2008; Steer, Ball, Ranieri, & Beck, 1999, Whisman, Perez, & Ramel, 2000). Association of the (mostly) affective items to either the cognitive or the somatic factor has remained unstable though.

However, some authors proposed complex factor models of the BDI-II (Arnau et al., 2001; Brouwer, Meijer, & Zevalkink, 2013; Ward, 2006), all of which implemented a bi-factor structure. The models comprised a general factor and two or three group factors, which were associated with cognitive and somatic items (Arnau et al., 2001; Brouwer et al., 2013; Ward, 2006) , and cognitive, somatic and affective items (Brouwer et al., 2013) respectively.

The mathematical similarity of the simple and complex factor models complicated resolving the dispute on the BDI-II factor structure via confirmatory factor analysis. Thus, studies that compared simple factor models with complex factor models obtained inconsistent findings. An extensive confirmatory factor analysis (CFA) study conducted by Vanheule, Desmet, Groenvynck, Rosseel and Fontaine (2008) examined the fit of ten different factor models of the BDI-II in two different samples. Four models revealed a better fit than the chosen reference model: three simple factor models (Buckley, Parker, & Heggie, 2001; Osman, Downs, Barrios, Kopper, Gutierrez, & Chiros, 1997; Viljoen, Grant, Griffiths, & Woodward, 2003) and one complex factor model (Ward, 2006). Another study (Quilty, Zhang, & Bagby, 2010) obtained good fit indices for the simple factor model by Osman et al. (1997) and the complex factor model by Ward (2006). A recently conducted CFA confirmed a good fit for the complex factor model by Ward (2006) and two additional complex factor models in a large sample of 1’530 outpatients with heterogeneous clinical syndromes (Brouwer et al., 2013). Thus, recent findings in the literature favor a complex factor model of the BDI-II over the simple factor models. Due to the good fit of the bi-factor model by Ward (2006), which was reported by many authors, we selected the model by Ward (2006) as a reference to test against the model presented in the current paper.

NMDS results of the BDI-II and activation in depression

The superiority of a complex factor structure of the BDI-II is supported by an analysis with Nonmetric Multidimensional Scaling (NMDS) which revealed an interpretable, dimensional item structure in the BDI-II (Bühler et al., 2012). Furthermore, Bühler et al. (2012) found an additional source of systematic variance, which was suggested to be related to the activation level of the BDI-II symptoms.

Activation looks back on a long history in the classification of depressive subtypes (e.g. Koukopoulos & Koukopoulos, 1999; Shorter, 2007). For example, the subtypes agitated and retarded depression, which were described by Klein and Davis (1969), have by definition a strong correlation to the activation level. Furthermore, these subtypes were also included in the influential Research Diagnostic Criteria (Spitzer, Endicott, & Robins, 1978). Well preceding these two manuscripts, Hamilton (1960) already proposed the two subtypes with respect to his factor analytic results of the Hamilton Depression Rating Scale. Another indication of an activation factor in the BDI-II has been found by Cohen (2008), who applied Multidimensional Scaling (of

which NMDS is a subtype) to the Beck et al. (1996) depression sample. He interpreted the solution as containing a fine grained matrix structure of 6 different categories of depressive symptoms times 3 categories of arousal. Although the solutions and the labels in the solution by Cohen (2008) and Bühler et al. (2012) slightly differ, the main results, indicating influences from cognitive, somatic and activation/arousal factors were confirmed. On the grounds of these findings it is astounding that none of the BDI-II factor models in the literature contained an activation factor. Thus, based on the results of the NMDS solution of the BDI-II by Bühler et al. (2012), the factor model presented in the current paper included an additional factor, which is argued to be related to the activation level of the symptoms.

Because NMDS is a rarely known method of analysis, the following subsection is intended to deliver an understanding of the key concepts in NMDS that are required to derive a factor model from an NMDS solution. A detailed mathematical description of NMDS can be found for example in Borg and Groenen (2005).

The representation of item structures in NMDS

NMDS allows to visualize the inherent structure of the data and to compare diverging factor structures. It can be applied to display the item structure of an inventory in a low dimensional space, usually of 2 dimensions (hereinafter referred to as symptom space), which has been used with great success to explore the symptom structures of psychopathological inventories (Bühler et al., 2012; Cohen, 2008; Läge, Egli, Riedel, & Möller, 2012; Steinmeyer & Möller, 1992). Within this symptom space, the pairwise similarities of any two symptoms are represented by the (Euclidean) distances between them. Hence, two highly similar symptoms are located in close distance to each other, whereas two dissimilar symptoms are located farther apart. To estimate the pairwise similarities between symptoms, the Pearson correlation coefficient can be used for example. Since the Pearson correlation coefficient is also applied in factor analyses, the resulting NMDS solution and the results from factor analyses yield highly similar results. Thus, from the postulated structure in a factor model, specific predictions can be obtained, how the postulated structure will be represented in an NMDS solution.

The representation of simple and complex factor structures in NMDS solutions largely diverge. In a simple factor structure, all pairwise item relations can be dichotomized into similar (pairings of items which are associated with the same factor) and dissimilar (pairings of items which are associated with different factors). In an NMDS solution, similar items are located close to each other while dissimilar items are located farther apart. Thus, similar items define distinct regions in the symptom space. Accordingly, items of simple factor structures are represented as multiple symptom clusters in NMDS solutions (left hand side of Figure 1). In contrast, in a complex factor structure, gradual differences in the items' similarities may arise due to different linear combinations of the associated factors. For example, items which load equally on two factors are as similar to items associated with the one as to items associated with the other

factor. These gradual similarity differences of the items allow for a dimensional item structure, which is reflected in the NMDS solution. Thus, items associated with multiple factors are located in a transition region between those items solely depending on their respective factors (right hand side of Figure 1).

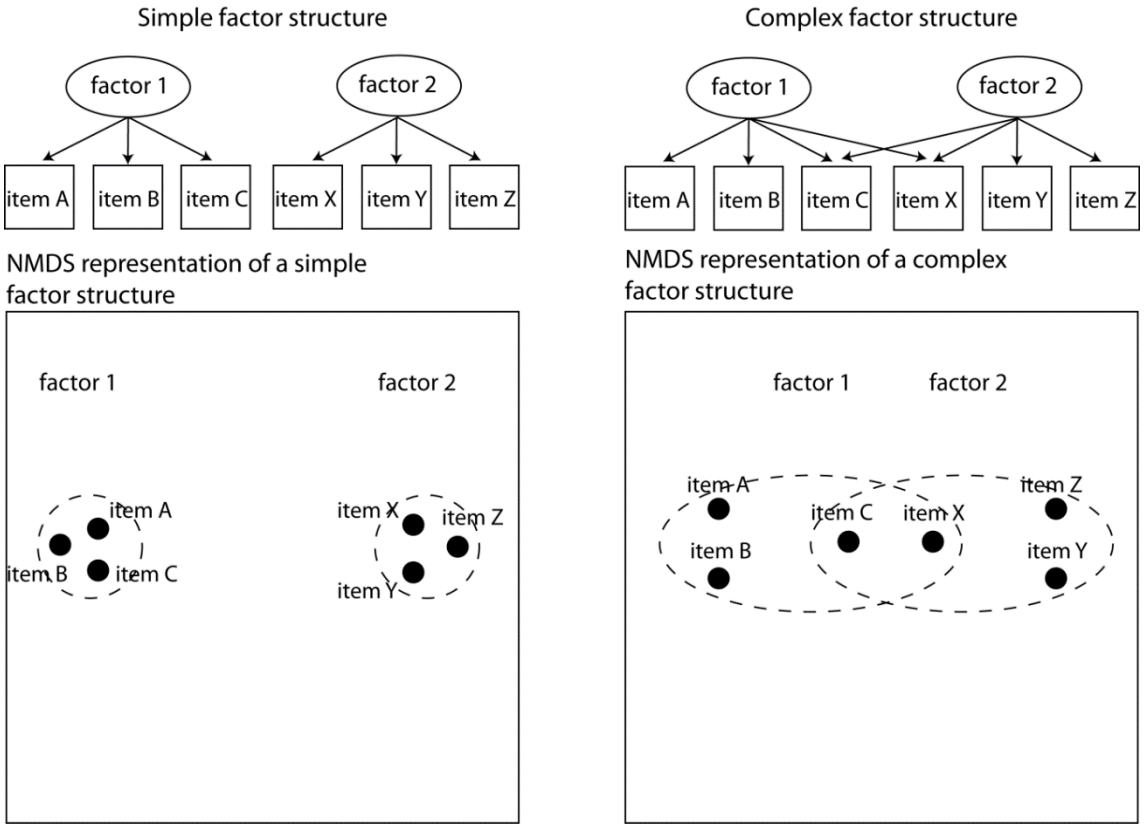

Figure 1. Representation of data structures in NMDS solutions. While the left hand side represents an NMDS solution of a simple factor structure, the right hand side represents an NMDS solution of a complex factor structure. In a simple factor structure, items A, B, C on the one hand and X, Y, Z on the other hand are considered to be measuring something similar, namely factor 1 and factor 2 respectively. Assume A loads high on factor 1 and B loads low on factor 1. They would still be considered to measure something similar because everything else the items measure must be considered error in a simple factor structure. In contrast, in a complex factor structure, items C and X are considered to measure something quite different, namely rather factor 1 and factor 2 respectively (even though they are associated with the exact same factors). Thus, complex factor structures allow for gradual differences in the similarity of two items, while simple factor structures do not. Instead, simple factor structures distinguish categorically between similar items (associated with the same factor) and different items (associated with different factors).

The degree of clustering in an NMDS solution can thus deliver hypotheses about the underlying item structure: a high degree of clustering indicates a categorical item structure (simple factor structure) while a low degree of clustering indicates a dimensional item structure (complex factor structure). Furthermore, the specific locations of items indicate the items' associations with underlying factors. On the right hand side of Figure 1, the items in the transition region between factor 1 and factor 2 indicate an association with both factors, moreover, item C is expected to yield higher loadings on factor 1 than factor 2 and, vice versa, item X is expected to yield higher loadings on factor 2 than factor 1.

When interpreting NMDS solution, one issue needs special attention: the relational approach of NMDS neglects a general factor (if inherent) in any item structure. NMDS solutions exclusively depend on object-interrelations and not on absolute values. Thus, a general factor is never directly reflected in an NMDS solution because it is associated with all items. However, items exclusively associated with a general factor still represent an item cluster in NMDS solutions because of their high similarity to each other. In the BDI-II, pronounced intercorrelations (e.g. Beck et al., 1996; Brouwer et al., 2013; Keller et al., 2008; Ward, 2006) have been reported. Any NMDS solution of the BDI-II should thus be interpreted with these previous findings in mind.

Based on the representations of simple and complex factor structures in NMDS solutions, a simple factor structure of the BDI-II would be indicated by distinct symptom clusters in the NMDS solution. Furthermore, the number of clusters in the NMDS solution should concur with the number of different factors. Additionally, no difference between orthogonal and oblique factor orientation are expected because highly intercorrelated items have no influence on the structure of NMDS solutions. In contrast, a complex factor structure would be indicated, if no obvious item clusters could be identified in the NMDS solution of the BDI-II. Instead, items associated with multiple factors would be expected to be located in transition regions between the clusters (if any existed) and thus define a pronounced dimensional symptom structure.

Adequacy of previous factor models and the derivation of a new factor model

The NMDS solution of the BDI-II (Figure 2) reported by Bühler et al. (2012) revealed a low degree of item clustering and thus indicated a complex factor structure. The cognitive items of the BDI-II were located mainly on the left hand side in Figure 2. However, they were spread over a broad region, almost obliterating a separation between cognitive and somatic items. The predominantly affective items, which were exclusively associated with the general factor in the bifactor models (Brouwer, 2013; Ward, 2006), were located in the center of the NMDS solution with only little scattering. However, there was only little space between the affective items and some of the cognitive and somatic items. Approaching the affective items from the right, the somatic items were even more scattered than the cognitive items, but were generally found to be located on the right hand side. The low degree of clustering in the NMDS solution of the BDI-II

indicates a better fit for the proposed bi-factor models (Brouwer, 2013; Ward, 2006) than for the simple factor models (e.g. Buckley et al., 2001; Osman et al., 1997; Viljoen, Grant, Griffiths, & Woodward, 2003). Consider for example the bi-factor model proposed by Ward (2006), which will be hereinafter referred to as the G2F model (a factor model with one general and two group factors; the model is depicted in Figure 3). The linear combination of the cognitive group factor and the general factor allows for the gradual location differences of the cognitive items: different loading patterns of the symptoms on the two factors may explain the symptoms' locations extending from the left to the center. Analogously, the linear combination of the somatic group factor and the general factor may account for the extension of the somatic symptoms from the right to the center. In contrast, a simple factor model would explain these gradual differences as random error.

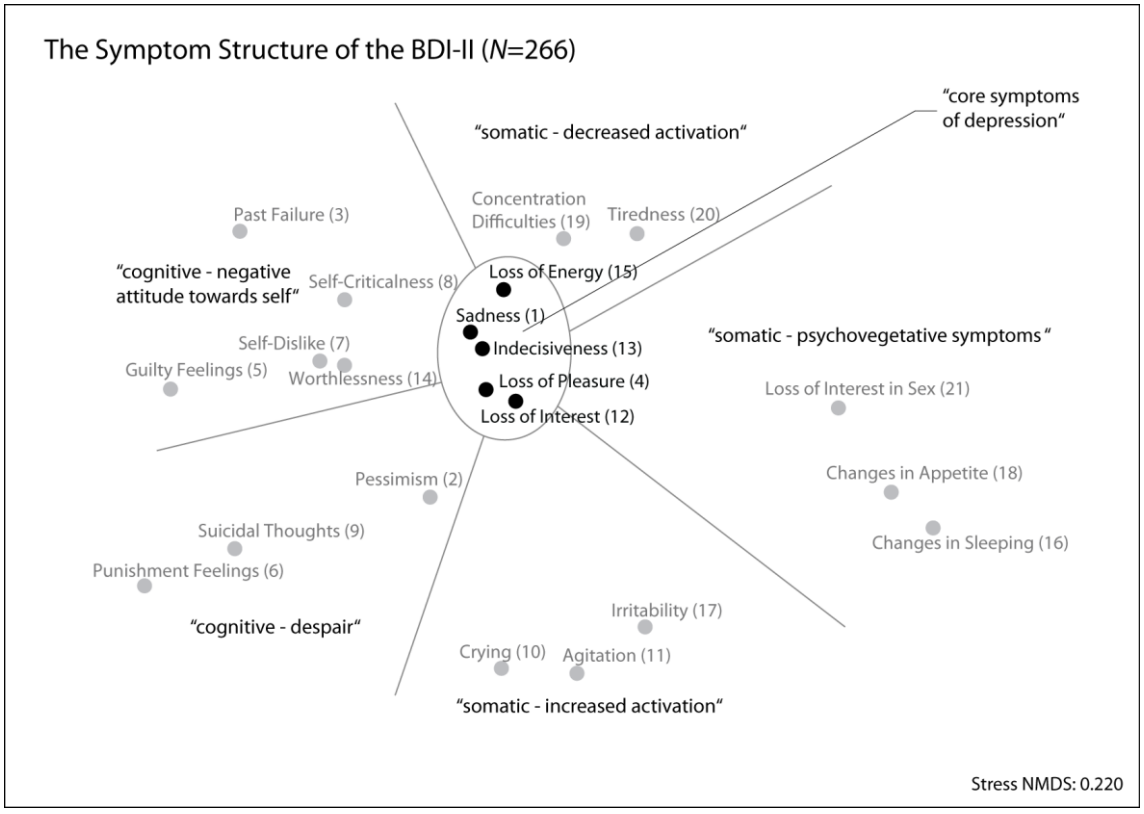

Figure 2. Qualitative aspects of BDI-II symptoms in the norming sample. The stress-value represents a badness of fit estimate for the NMDS solution. Adapted from "Die Symptomstruktur des BDI-II: Kernsymptome und qualitative Facetten. [The symptom structure of the BDI-II: core symptoms and qualitative facets]" by J. Bühler, F. Keller and D. Läge, 2012, *Zeitschrift für Klinische Psychologie und Psychotherapie*, 41(4), p. 240. Copyright by Hogrefe. Adapted with permission.

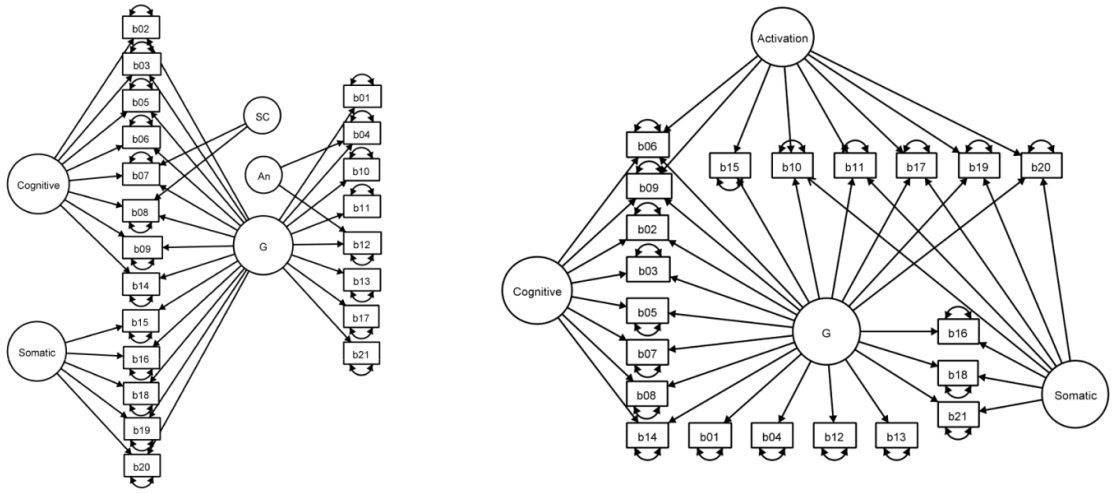

Figure 3. Path diagrams of the G2F model (left hand side) and the G3F model (right hand side). SC = Self-Criticism, An = Anhedonia, G = General Factor.

However, besides the variability introduced by a cognitive, a somatic and a general factor, the NMDS solution of the BDI-II revealed an additional source of interpretable variance. Bühler et al. (2012) proposed a more detailed structure of the symptoms by defining six distinct regions (aspects of depression) labeled "negative attitudes towards self" and "despair" (both associated with a cognitive factor), and "decreased activation", "psychovegetative symptoms" and "increased activation" (associated with a somatic factor). The sixth region, which comprised mostly affective items, was labeled "core symptoms of depression". As the labels in the NMDS solution indicate, Bühler et al. (2012) suggested differentiating between five sources of systematic variance. The first and second source of variance were proposed as the two commonly replicated somatic and cognitive factors, whose associated items were located on the right and on the left hand side of the NMDS solution respectively. Thus, it was argued that these two factors accounted for the variability of the items location in the horizontal. The third source of variance was proposed as a general factor, which comprised all items, but with an emphasis on the (mostly) affective items which constituted the core symptoms of depression. The fourth source of variance was suggested to constitute a factor related to the activation level of the symptoms. Its associated items were located in the upper and the lower part of the NMDS solution. The fifth source of variance was suggested to differentiate between "despair" and "negative attitude towards self". However, this last source of variance was neglected in the following factor model to ensure uniqueness of the items' locations and to apply a more parsimonious model. Following the hypothesis that indeed all the locations of the symptoms in the NMDS solution of the BDI-II are governed by the first four sources of systematic variance, we derived the specific linear combinations of the factors directly from Figure 2. The resulting factor model is given in Figure 3.

Additionally, the symptoms' locations in the NMDS solution allowed specific hypotheses about their expected factor loadings. Thus, higher loadings on the cognitive factor were expected for those items in the periphery on the left hand side (items 3, 5, 6, and 9) than for those items closer to the center (Items 2, 8, 7, and 14). Vice versa, the items 2, 8, 7, and 14 were expected to load higher on the general factor than the items 3, 5, 6, and 9. Likewise, the peripheral items of the somatic factor on the right hand side, items 16, 18 and 21, were expected to load higher on the somatic factor than the items 20, 17, 19, 11 and 10 and, again, vice versa to reveal lower and higher loadings on the general factor respectively. The linear combinations of these two factors and the general factor were expected to explain the variability of the items' locations in the horizontal of Figure 2. Please note: if the underlying structure followed a simple factor structure instead, the specific locations of the items at the left (cognitive) and at the right (somatic) were completely random, except that the cognitive items would still be left and the somatic items would still be right.

The labels increased and decreased activation of the aspects in the NMDS solution imply one single factor (an activation factor) instead of two separate factors to explain the upper and the lower item locations in Figure 2. Thus, items at the top of Figure 2 were expected to yield loadings with a reversed sign on the activation factor compared to the items at the bottom. Please note: if the source of variation was solely governed by random error (i.e. if the activation level had no systematic influence), the sign of the loadings would vary randomly.

The specifications of the factor model were very specific at this point already. Of course, theoretical considerations should not be neglected. Hence, a few adjustments to the model were made on theoretical grounds. Firstly, "past failure" was not considered to be associated with the activation factor even though its upper location in the NMDS solution would suggest so. Secondly, "loss of energy" was considered to be associated with the activation factor, even though it was categorized as a core symptom in the NMDS solution. With these two adjustments, a theoretically sound factor model was retrieved which is depicted in Figure 3 and was drawn with the program Onyx (von Oertzen, Brandmaier, & Tsang, 2012). In contrast to the previous complex factor models in the literature, which exclusively implemented a bi-factor structure, the herein presented factor model deviates from a bi-factor structure. Specifically, the deviations concern those items that yield multiple dependencies on the "group" factors (items 6, 9, 10, 11, 17, 19, and 20). To simplify referring to the proposed factor model and to contrast it with the G2F model, the model was labeled G3F (a factor model with one general and three additional factors; Figure 3).

Our second aim, which constitutes the remainder of the manuscript, was to assess the goodness of fit of the G3F model and to compare it with the fit of the G2F model. The fit of the two models was assessed in two different samples. Firstly, the sample that was applied in the NMDS analysis by Bühler et al. (2012) was used to ascertain the adequacy of the G3F model

(i.e. the adequacy of the derivation of the model). Secondly, an additional, independent set of data was used to confirm the goodness of fit of the G3F model and to compare it with the fit of the G2F model.

Methods

Participants

Two samples were used to assess the goodness of fit estimates of the factor models. The first sample was the norming sample of the German version of the BDI-II (Hautzinger, Keller, & Kühner, 2006). It consisted of depressive in- and outpatients (N = 266), whose BDI-II data were collected within the normal course of treatment at different counseling centers and psychiatric hospitals across Germany. The mean age of the sample was 48.8 years (SD = 15.7) and 65.4% of the patients were female. The mean BDI-II total score was 23.4 (SD = 13.0).

The replication sample consisted of patients from a clinic for psychosomatic disorders (N = 898); they completed the BDI-II at admission within the routine diagnostic procedure. Only those patients diagnosed with a primary affective disorder (ICD-10: chapter F3x) as their main diagnosis were included in the analysis. This resulted in a reduced set of 569 patients. The mean age in the replication sample was 46.7 years (SD = 8.9) and 64.1% of the patients were female. The mean BDI-II total score was 24.1 (SD = 10.6). The three most frequent comorbid disorders were substance related disorders (ICD-10: F1; n = 113 (19.9%)), somatoform disorders (ICD-10: F45; n = 104 (18.3%)), and anxiety related disorders (ICD-10: F40 & F41; n = 92 (16.2%)).

Procedures

All analyses were conducted with MPlus 7 (Muthén & Muthén, 2012). To account for nonmetric data, the WLSMV estimator for ordered categorical (ordinal) data was applied. The WLSMV adjusts the means and variances of the test-statistics to approximate chi-square distributions (Asparouhov & Muthén, 2010). The models were evaluated using the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA). A CFI $\geq$ .95, an RMSEA value $\leq$ .06, and a TLI $\geq$ 0.95 are considered a good fit, following the guidelines of Hu and Bentler (1999). A reasonable fit is indicated for values of CFI $\geq$ .90, and RMSEA $\leq$ .08 (Kline, 2005; Browne & Cudeck, 1993). For the G2F model, error covariances were included for the items “Loss of Pleasure (4)” and “Loss of Interest (12)”, and “Self-Dislike (7)” and “Self-Criticalness (8)” as described by Ward (2006) and depicted in Figure 3. For the G3F model, no error covariances were assumed. The factors in both models were specified as being orthogonal to each other. Thus, the correlations between factors in the models were set to zero.

Results

In the norming sample, from which the NMDS solution in Figure 2 was also obtained, a CFA revealed good fit indices for both models, the G2F and the G3F (Table 1). Judging only by the value of the indices, the G3F model revealed slightly better fit indices than the G2F model in all three fit indices though. The CFA based on the replication sample obtained an acceptable RMSEA and good CFI and TLI values for the G2F model (Table 1). For the G3F model, goodness-of-fit increased substantially, revealing a good fit for all three measures (Table 1). In the following, detailed results are only given for the replication sample (N=569), because the G3F model was derived from an NMDS solution based on the norming sample and thus may be positively biased.

In the replication sample, the G2F model revealed one factor loading that did not reach significance - "Changes in Appetite" only showed a small loading (0.10) on the somatic factor. Additionally, the postulated covariance between "Self-Dislike" and "Self-Criticalness" was found to be non-significant (-0.08). In the G3F model, "Crying" failed to load significantly on the activation factor and on the somatic factor with loadings of -.09 and .08, respectively. Additionally, "Irritability" failed to load significantly on the activation factor. All other items were found to load significantly on the postulated factors (Table 2).

The majority of the predictions, which were postulated in the previous section, were found to hold. With respect to the activation factor, the sign of the item loadings were reversed of those items in the upper compared to the items in the lower part of Figure 2. Thus, random locations with respect to the vertical order of the items in the NMDS solution are very unlikely. However, the activation factor was mainly defined by the common variance of the items associated with a low level of activation (items 15, 19, and 20), which was indicated by the lower absolute values of those items associated with a high level of activation (items 6, 9, 10, 11, and 17).

Table 1

Fit indices of the G2F and the G3F model in the norming and the replication sample

Model	χ^2	*df*	CFI	TLI	RMSEA
norming sample (N=266)					
G2F (Figure 3; Ward, 2006)	268.2	174	0.986	0.983	0.045
G3F (Figure 3)	216.6	165	0.992	0.990	0.034
replication sample (N=569)					
G2F (Figure 3; Ward, 2006)	541.8	174	0.962	0.954	0.061
G3F (Figure 3)	371.0	165	0.979	0.973	0.047

Table 2

Factor loadings of the G3F model in the replication sample of 569 patients diagnosed with a primary affective disorder

BDI-II item	G-factor	Cognitive	Somatic	Activation
Sadness (1)	0.69			
Pessimism (2)	0.68	0.14		
Past Failure (3)	0.46	0.60		
Loss of Pleasure (4)	0.82			
Guilty Feelings (5)	0.48	0.57		
Punishment Feelings (6)	0.47	0.45		-0.16
Self-Dislike (7)	0.57	0.46		
Self-Criticalness (8)	0.59	0.49		
Suicidal Thoughts (9)	0.44	0.24		-0.20
Crying (10)	0.57		0.08^{n}	-0.09^{n}
Agitation (11)	0.47		0.43	-0.17
Loss of Interest (12)	0.83			
Indecisiveness (13)	0.76			
Worthlessness (14)	0.66	0.43		
Loss of Energy (15)	0.76			0.43
Changes in Sleeping (16)	0.21		0.47	
Irritability (17)	0.46		0.54	-0.03^{n}
Changes in Appetite (18)	0.31		0.39	
Concentration Difficulties (19)	0.67		0.19	0.31
Tiredness (20)	0.60		0.32	0.44
Loss of Interest in Sex (21)	0.46		0.16	

Note. ^{n} indicates a non-significant loading.

The values of the factor loadings revealed by the cognitive items followed the predicted order closely: the peripheral items “Past Failure” and “Guilty Feelings” revealed higher factor loadings than the more central items “Self-Dislike”, “Self-Criticalness” and “Worthlessness” (Figure 2). Also, “Punishment Feelings” revealed a higher loading than “Suicidal Thoughts”, and “Suicidal Thoughts” revealed a higher loading than “Pessimism” (Figure 2). So far, the order of the factor loadings was in perfect accordance with the predictions derived from the NMDS analysis. However, the items of the aspect “despair” revealed over all reduced factor loadings compared to the items of the aspect “negative attitude towards self”.

The factor loadings of the somatic factor showed a less concordant image: Indeed, the peripheral items “Changes in Sleeping”, and “Changes in Appetite” yielded amongst the highest factor loadings, but so did the more central items “Irritability”, “Agitation” and “Tiredness”. Furthermore, the small loading of the item “Loss of Interest in Sex” is noteworthy, despite its location in the periphery and its close distance to the other psychovegetative items “changes in appetite” and “changes in sleeping” in Figure 2.

The factor loadings on the general factor (G-factor) were in good accordance with the predictions. The loadings on the G decreased in accordance with increasing distance from the center of the NMDS solution (Figure 2), except for the item "Crying". Also, the items "Punishment Feelings" and "Suicidal Thoughts" revealed similar loadings despite their different distances from the center.

Discussion

Our aims were (a) to present a new factor model for the BDI-II and (b) to demonstrate a procedure, how factor models can be derived from NMDS solutions. The G3F model (Figure 3) was derived from a previously published NMDS solution of the BDI-II (Bühler et al., 2012), which indicated a complex factor structure with an additional factor related to the activation level of the symptoms. Thus, the G3F model included four factors, a G-factor, a cognitive, a somatic, and an activation factor. It did not integrate a strict bi-factor structure, which allowed for some items to load on multiple factors besides the G-factor. However, adjustments in item associations to the cognitive and somatic factors were only small compared with the G2F model. Nevertheless, the conducted CFA, which was based on an independent data set, indicated substantially better fit indices for the G3F model than for the G2F model.

As others have argued before us (e.g. Brouwer, et al. 2013; Quilty et al., 2010), complex factor models (bi-factor models) have been found to represent the BDI-II item structure more adequately than simple factor models. Thus, only the G2F model and none of the simple factor models were tested against the G3F model in the current study.

The additional activation factor that was included in the G3F model is especially promising regarding the theoretical foundation of depression: the classification of depressive subtypes according to specific activation levels looks back on a long history in depression research, and the concept has been applied time and again (Cohen, 2008; Hamilton, 1960; Klein and Davis, 1969; Koukopoulos & Koukopoulos, 1999; Shorter, 2007; Spitzer et al., 1978). It astounds all the more that none of the common BDI-II factor models included an activation factor, even though some items strongly suggest its existence (e.g. "Agitation", "Loss of Energy").

We agree with Brouwer et al. (2013) and other authors before them (e.g. Quilty et al., 2010; Ward, 2006) that the general factor represents the main component in the BDI-II symptom structure and that it should best be interpreted as a factor of depression severity. However, we believe that the additional factors (cognitive, somatic, and activation) should not be neglected either with respect to a scientific discourse on depression. There are two capital reasons why these factors are important for the construct of depression.

Firstly, we found substantial loadings of the items on their respective factors – even after the G-factor was accounted for. While the (fisher transform corrected) mean loading on the G-factor amounts to $\bar{r}_G = .59$, the mean loading on the remaining factors still yielded $\bar{r}_C = .43$, $\bar{r}_S = .33$, and $\bar{r}_A = .23$ for the cognitive, the somatic, and the activation factors respectively. Not to include these factors would mean to discard reliable information about systematic differences in the symptom profiles of depressive patients. Thus ultimately, we came to a different conclusion than Brouwer et al. (2013), who denied the importance of additional factors in the BDI-II. This interpretational difference may be explained by differences in the sample (our sample consisted of patients with an affective disorder only, whereas Brouwer et al. (2013) examined a diagnostically mixed sample) and by differences in the calculation of the factor importance (ECV vs. mean loadings). However, in our view, the ECV (i.e. factor variance in proportion to common variance of all factors) should not be the measure of choice for factor importance, if factors with quite different numbers of items were compared – especially in an inventory that has been proven to yield high internal consistency. For the ECV, there is no "penalty" for low item loadings. Instead, each additional item on a factor improves the ECV value of that factor and thus biases the measure towards factors with many items. In contrast, the mean item loading corrects for this dependency on the number of items.

The second reason aims for a broader understanding of our depressive patients' BDI-II data: a factor model of the BDI-II items is more than a simple representation of the structure of this specific questionnaire. In case the BDI-II is applied to a representative sample of depressive patients, it is a representation of the symptom structure of depression itself (given by the framework of the DSM-IV). As a result, structural models of the BDI-II can (and should) be used to gain new insights into depression. In this respect, we believe it is wise to consider any systematic pattern in its item structure as meaningful; especially in the field of depression, in which many authors question the homogeneity of the disorder (e.g. Fink & Taylor, 2007; Joiner, Walker, Pettit, Perez & Cukrowicz, 2005; Lichtenberg & Belmaker, 2010; Parker, 2007; Shorter, 2007; Stewart, McGrath, Quitkin & Klein, 2007). A fine grained model of depressive symptoms may yield great implications; for example to disentangle the response rates of different subgroups to specific treatments.

In the current study we were able to show that the factors in the G3F model of the BDI-II are indeed structurally stable and that the derivation of factor models from NMDS solutions constitutes a potent procedure to obtain empirically supported factor models. Based on previous NMDS results (Bühler et al., 2012), the expectations about the structure and the (rank) order of the factor loadings in the G3F model were confirmed by an independent set of data. However, there was one noteworthy exception to the generally good accordance of the model. The item "Crying" did not reveal significant factor loadings on any of the postulated additional factors in the replication sample (instead it showed slightly increased loadings on the general factor to

what was expected), though it loaded significantly on the activation factor in the norming sample ($\lambda = -0.28$; the results of the norming sample were not shown in the current article). Thus, the different loading patterns of the item "Crying" in the two samples suggest that the item's loadings were rather unstable. Furthermore, the communality of the item "Crying" in the replication sample was amongst the lowest communalities obtained, suggesting a rather idiosyncratic relevance within the BDI-II. Similarly, low communalities for this item were obtained in the five data sets analysed by Ward (2006); and in the Quilty et al. (2010) study, the loadings of "Crying" attained amongst the lowest scores on the explaining factors. Furthermore, difficulties with the items' response categories have been reported when item response models were applied: Beck et al. (1996) noted that the categories did not display the anticipated rank order. Additionally, Hautzinger et al. (2006) reported that the categories discriminated insufficiently and that the category thresholds were partially wrongly ordered.

There are some limitations to the current study though. Both patient samples were collected in Germany and in both studies the German version of the BDI-II was applied. Even though the German version of the BDI-II was carefully adapted to be in line with the English version (Hautzinger et al. 2006), systematic differences in the covariance structure cannot be ruled out completely. Furthermore, not all items loaded on the activation factor as high as expected. Even though the two groups of items (with assumed low and high activation levels) could be identified by their sign on the activation factor, items with a high level of activation did not reveal loadings as pronounced as items with a low level of activation. Two of the items with an assumed high level of activation did not reveal a significant loading on the factor at all: "Irritability" and "Crying". The results suggest that these items, at the bottom of Figure 2, are not necessarily associated with an activation factor in both samples. With respect to the multidimensional scaling solution by Cohen (2008), an association of "Irritability" with the activation factor seems likely (Irritability was found by Cohen (2008) to be associated with arousal) and its lack of association with the activation factor in the replication sample may merely be sample specific. However, the item "Crying" was not associated with the arousal dimension found by Cohen (2008). Thus, the item's vertical location in Figure 2 may indeed be governed by random error in the NMDS solution by Bühler et al. (2012).

In the current article, we demonstrated the link between factor models and NMDS solutions in the analysis of questionnaire data. Hence, given the symmetry in the two methods, the question may arise: Do we need different methods to assess the item structure of an inventory? Obviously, if multiple methods are applied to assess the same object of investigation, methodological artifacts are minimized. However, beyond the argument of methodological artifacts, the methods yield specific benefits for future studies. In this regard, we believe that the answer is twofold. The structure of an inventory might be easier to handle as a factor model. Factor models are far more common, they are easier to impose constraints on and inferential statistics are applicable. Then again, we believe that the visual representation of the symptom structure in the

NMDS solution helps to achieve a better understanding of the empirical covariance structure and thus imposes an excellent framework to postulate hypotheses about structural patterns, as we have successfully demonstrated with the BDI-II symptom structure.

Our analysis in this article suggests that the symptom structure in the BDI-II is quite stable, even as a fine grained dimensional structure. Thus, we are confident that the structure of the BDI-II indeed comprises multiple factors, which may at the least be beneficial to further enhance our knowledge about depression and, eventually, to develop more specific treatments for our depressive patients.

Acknowledgements

We thank Dr. Robert Mestel, head of Research/Quality Assurance of HELIOS Klinik Bad Grönenbach, for providing us with the dataset of the replication sample.

The Symptom Structure of the Hamilton Depression Scale (HAM-D): an NMDS analysis

Joël Bühler[1], Florian Seemüller[2], Damian Läge[1]

[1]Department of Psychology, University of Zurich, Zurich, Switzerland
[2]Departement of Psychiatry and Psychotherapy, Ludwig-Maximilians-University Munich, Munich, Germany

Submission status:

Submitted to *The British Journal of Psychiatry*, August 2013.

Authors' contributions:

Joël Bühler: Development of the research question, review of the literature, execution of the analyses, interpretation of the results, writing of the manuscript

Florian Seemüller: Provision of the data, revision of the manuscript

Damian Läge: Development of the research question, supervision and discussion of Joël Bühler's contributions, revision of the manuscript

Abstract

Background: The Hamilton Depression Rating Scale (HAM-D) is heavily debated in the literature. The debate has focused the diverging results of factor analyses in the literature mainly which raised concern about the validity and the general usefulness of the scale.

Aims: To re-analyse the HAM-D with a different methodological approach and thus to integrate the diverging factor analytic results.

Method: Nonmetric Multidimensional Scaling (NMDS) was applied to two sets of data. One set of data was collected in a large prospective, naturalistic, multicentre study and consisted of 1073 patients. The other set of data were previous factor analytic findings which were reanalysed via NMDS.

Results: Both NMDS solutions yielded very similar results. Four theoretically sound groups of symptoms could be identified in both data sets and the symptom-groups revealed a similar dimensional structure in both sets. The proposed dimensional symptom structure was found to be in accordance with theoretical considerations.

Conclusions: Most previous analyses modelled the HAM-D item structure as simple factor structures. However, both the current findings suggest that the structure of the HAM-D items is rather dimensional which cannot be properly modelled with simple factor structures. Thus, the current results may explain the diverging findings of previous factor analyses.

Key words: depression, Psychiatric Status Rating Scales, Evaluation Studies

Introduction

The Hamilton Depression Rating Scale (HAM-D) is the most widely used clinician-administered rating scale for depression (Williams et al., 2008) and is still considered the "gold standard" in efficacy evaluations of antidepressant drugs (Helmreich et al., 2012). However, many authors have pointed out flaws in the scale; most frequently its multidimensional item structure, its moderate sensitivity to change and some specific items were criticized as measuring multiple constructs with one single item (e.g. Bagby, Ryder, Schuller, & Marshall, 2004; Faries, Herrera, Rayamajhi, DeBrota, Demitrack, & Potter, 2000; Pancheri, Picardi, Pasquini, Gaetano, & Biondi, 2002; Santen, Gomeni, Danhof, & della Pasqua, 2008; Santor & Coyne, 2001).

To overcome the issue of multidimensionality and sensitivity to change of the original HAM-D, some authors constructed unidimensional subscales: the HAM-D_6 (Bech, Gram, Dein, Jacobson, Vitger, & Bolwig, 1975) the Maier and Philip Severity subscale (Maier & Philipp, 1985), the Gibbons Global Depression Severity subscale (Gibbons, Clark, & Kupfer, 1993) and the Toronto HAM-D_7 (McIntyre, Kennedy, Bagby, & Bakish, 2002). Moreover, two subsets were described by Evans, Sills, DeBrota, Gelwicks, Engelhardt, & Santor (2004) and by Santen et al. (2008) on the basis of item characteristics. All the subscales included six to eight items which were drawn from a reduced set of only 12 HAM-D items. Therefore, the scales overlap substantially. An overview of item inclusion in the different scales can be obtained from Table 1.

Unidimensionality of summed symptom severity is a necessary prerequisite when detecting change. If a measure would consist of several different dimensions and a lack of change was observed in the total, it could not be determined if this was attributable to constancy or to shifts in symptomatology on the different dimensions. The progress on a dimension of interest may be even masked by change on another dimension of lesser interest. Thus, it is possible, as Bagby et al. (2004) noted that positive treatment effects may had been underestimated due to the impact of side effects on a (for example) somatic dimension. This paradigm has been heavily debated in the last two decades, also in regards to the comparability of different classes of antidepressants. For example Möller (2001) assumed reduced sensitivity of the HAM-D to SSRI antidepressants compared to the older tricyclic antidepressants, because some of the side effects of SSRIs (e.g. sleep disturbances, agitation and gastrointestinal symptoms) may affect the according HAM-D items.

Table 1

The unidimensional subscales of the HAM-D items

HAM-D-Item	Bech et al. (1975)	Gibbons et al. (1993)	Maier & Philipp (1985)	Evans (2004)	Santen et al. (2008)	Toronto (McIntyre et al., 2002)
1. depressed mood	X	X	X	X	X^1, X^2	X
2. guilt	X	X	X	X	X^1, X^2	X
3. suicide		X			X^1	X
4. insomnia, initial						
5. insomnia, middle					X^2	
6. insomnia, late					X^2	
7. work and interests	X	X	X	X	X^1, X^2	X
8. retardation	X		X		X^1	
9. agitation		X	X			
10. anxiety, psychic	X	X	X	X	X^1, X^2	X
11. anxiety, somatic		X		X		X
12. somatic, gastroin-testinal						
13. somatic, general	X			X	X^1, X^2	X
14. genital symptoms		X				
15. hypochondriasis						
16. weight loss						
17. insight						
18. diurnal variation						
19. depersonalization / derealization						
20. paranoid symp-toms						
21. obsessional symp-toms						

Note. [1] and [2] refer to the two different subsets in the study by Santen et al. (2008) based on different selection criteria.

Reducing the set of items, however, introduces serious drawbacks also. As Zimmerman, Posternak, and Chelminski (2005) pointed out, one of the major advantages of the HAM-D is the assessment of associated symptoms of depression besides the assessment of depressive core symptoms because associated symptoms are commonly used to select antidepressants (Zimmerman et al., 2004). Moreover, Bech et al. (1981) mentioned that the HAM-D fulfilled other purposes besides the assessment of severity, e.g. looking for item profiles. For this purpose, the scales full breadth is inevitable. Furthermore, subscales cannot cover all aspects of depression, although they might be important for a comprehensive understanding of the disease (Helmreich et al., 2012).

Hence, despite the heavy critique that has been expressed lately, the HAM-D must have provided useful information, or else it would not have withstood a decade of psychiatric and psychological research and two major revisions of the DSM manual. It is most likely not only its sum-score (because the rarely used short version would satisfy this need much better) but rather its capability to assess depressive symptomatology quite exhaustively that accounts for its widespread use – even though some items may not be unidimensional or yield the same importance for depression. Thus it may be exactly the HAM-D's breadth of depressive symptoms that makes it such a useful tool for exploring the structure of depression.

A solid knowledge about the relations of depressive symptoms is of vital importance not only for a coherent theory of depression, it yields practical implications as well; for example to disentangle the response rates of different subgroups to specific treatments (e.g. Baumeister & Parker, 2012; Lichtenberg & Belmaker, 2010). Even though the HAM-D does not fully cover the depressive syndrome as defined in DSM-IV (Möller, 2001), it still gives a comprehensive overview of depressive and associated symptoms. Hence, understanding the item structure of the HAM-D grants knowledge on the structure of depression itself and may help to promote new insights into the disease. Despite a substantial number of studies on the factorial structure of the HAM-D, a consensus about its structure could not be achieved (Bagby et al., 2004). Even the number of underlying factors found in the studies varied substantially from two (Steinmeyer & Möller, 1992) to eight (O'Brien & Glaudin, 1988; von Giesen, Bäcker, Hefter, & Arendt, 2001).

This vast variability in the number of factors suggests a highly unstable factor structure behind the HAM-D items. More precisely, it suggests an unstable *factorial simple* structure (i.e. an item to factor allocation in which each item is at the most allocated to one factor) behind a large part of the items. Despite the variability of the findings, some items were coherently allocated to the same factors. For example, most studies constituted a separate factor of insomniac symptoms; hence, it seems safe to assume an underlying sleep related factor (Bagby et al., 2004). However, the remaining symptoms loaded on different factors in the respective studies, and thus prevented a consistent factor allocation. Diverging results in factor analyses do not necessarily indicate diverging item structures though. If the underlying item structure was to follow a *factorial complex* structure (a structure in which items may be allocated to more than one factor), the results of exploratory factor analyses (EFA) need not necessarily be homogenous (because EFAs generally feature simple factor structures to avoid the indeterminacy problem). A slight difference in the data may cause even substantially diverging item to factor allocation because of the categorical approach (only one factor per item) and the data driven rotation of factors inherent to EFAs.

Cole et al. (2004) conducted a confirmatory factor analysis (CFA) of the HAM-D items for their own rationally derived model and for selected models of previous findings in EFA

studies. However, they chose a factorial simple structure for their Cole & Motivala Model (Cole et al., 2004). Although complex factor structures could easily be modelled in CFA, the Cole & Motivala Model did not integrate multifactor dependencies on any of the HAM-D items, which may have prevented the model from reaching adequate fit indices. They concluded that the structural models of the HAM-D still need further refinements instead. Evidently, the desired insights cannot be retrieved from yet another factor analysis study with yet another set of data. Nonmetric Multidimensional Analysis (NMDS) proposes an excellent tool to compare and integrate diverging factor analytic results as Bühler, Keller, and Läge (2012) demonstrated for the BDI-II. NMDS combines categorical and dimensional features and thus overcomes the flaws of exploratory factor analytic studies. NMDS is not a new method to investigate the symptom structure of the HAM-D either. An NMDS solution of the HAM-D_{17} by Steinmeyer and Möller (1992) has provided substantial indication that the underlying item structure does not follow a factorial simple structure. Even though NMDS results are very detailed in regards to the underlying item structure they are limited in generalization. Hence, the results of NMDS analyses should be replicated with independent data. Moreover, to increase generality of the results HAM-D_{21} and HAM-D_{17} should be examined. The analysis presented below covers all these aspects.

Aim of the study

The aim of the study was twofold. Firstly, we examined the item structure of the full HAM-D_{21} from a large dataset of depressive patients with Nonmetric Multidimensional Scaling (NMDS). To maximize generalization of our HAM-D_{21} NMDS analysis, a second NMDS analysis was conducted with a different set of data. This second analysis was based on previous findings of the HAM- D_{17} factor structure: co-occurrences on the same factors in previous studies were treated as similarity data and analysed via NMDS.

Methods

Sample characteristics

The sample was collected in a large prospective, naturalistic, multicentre follow-up study funded by the German Federal Ministry of Education and Research (BMBF). Subjects were recruited at six German psychiatric university hospitals and three psychiatric district hospitals. Inclusion criteria required age between 18 and 65 and signed informed consent. Patients had to meet diagnostic criteria according to ICD-10 (World Health Organization, 1992) for any major depressive episode (ICD-10: F31.3x–5x, F32, F33) or for depressive disorder not otherwise specified (ICD-10: F34, F38, F39), including bipolar depression (F31.2-3) melancholic depression (F32.9) and depression with psychotic symptoms (F32.2). Moreover, the diagnosis was confirmed by the Structured Diagnostic Interview of DSM-IV (SCID) (Wittchen, Wunderlich,

Gruschwitz, & Zaudig, 1997) and a distinction between bipolar I and bipolar II disorder based on DSM-IV criteria was made. A sample of 1073 Patients had been recruited and the patients were tested with the 21-items HAM-D scale (Collegium Internationale Psychiatriae Scalarum, 1977). Only the first rating, obtained before onset of treatment, was included in the current analysis. Also, to reduce possible effects of structural heterogeneity of the HAM-D symptoms in regards to disorder characteristics, only patients with unipolar depressive disorders were included in the final analysis (F32 & F33). The combined inclusion criteria resulted in a reduced data set which consisted of 911 patients. Of these 911 patients included, 62.8% were female and 37.2% were male. Their age varied about a mean of $M = 44.5$ with a standard deviation of $SD = 12.0$. The three most frequent comorbid diagnoses were dysthymia (F34.1: 50 cases), anxious personality disorder (F60.6: 39 cases) and agoraphobia (F40.0: 27 cases). All of the 911 HAM-D ratings were fully completed. No missing data was present in the dataset.

Procedures

Nonmetric Multidimensional Scaling was used to analyse the data. Both NMDS analyses were conducted with the software package Prodax (Oberholzer, Egloff, Ryf, & Läge, 2008). The final NMDS solution was calculated via bootstrapping from the original set of data to improve stability of the results (Bühler & Läge, 2013). It is worth mentioning that deviations of the results between the single step NMDS solution and the bootstrap NMDS solution usually are small.

NMDS transforms similarity relations – such as correlations or co-occurrence data – between objects into (Euclidean) distances and maps these distances into an *n*-dimensional space (Borg & Groenen, 2005). The transformation of similarity data into distances follows a rank ordering principle: The smaller the similarity between two objects is the larger the resulting distance to each other becomes. The mapping of these distances into *n*-dimensional space is attempted in the lowest dimensionality possible. In most cases, a two dimensional space is already sufficient for an adequate reproduction of the data structure. However, a perfect match between theoretically obtained distances (ideal distances) and distances measured in NMDS space (real distances) is hardly ever achieved. Hence, NMDS uses an iterative algorithm to determine the best possible fit between ideal distances and real distances of objects to each other. The deviation between ideal distances and real distances is measured as stress and indicates the difference in the rank order of the distances compared to the rank order of the similarities.

Two dimensional NMDS solutions resemble bi-plots of principal components analysis (PCA) at a first glance. However, this resemblance is misleading. In two dimensional PCAs the data are projected on a plane, which is spanned by those (data inherent) dimensions that are best able to explain the variance in the data. This simple projection completely disregards the variability on the other dimensions (which explain the residual variance) though. Contrarily in NMDS, the structure is modelled with respect to the full information in the covariance matrix. The solution is a distance structure (a distances matrix) in two dimensional Euclidian space.

Usually, these spaces (the PCA bi-plot and the NMDS solution) share some common features because obviously, the principal components of the variance between the items largely influence the NMDS solution as well. The additional information available in NMDS accounts for better item specific estimates with respect to the overall structure of the data though.

NMDS solutions directly reflect the results from factor analyses under certain conditions. The prerequisites and key-ideas to compare NMDS solutions and factor analyses results will be described in detail in the following paragraphs. A first prerequisite demands that the analysed coefficients be the same. Thus, we used Pearson correlation coefficients as similarity data to calculate the reported NMDS solution of our first analysis. Second, more a reminder than an actual prerequisite: NMDS models the variance within the (relational) similarity matrix only. Hence, a general factor (or any higher order factors) will not be displayed. When interpreting an NMDS solution, one must keep this purely relational model in mind.

We hypothesized in the introduction that the diverging factor analytic results of the HAM-D may be due to a *factorial complex* structure. A *factorial complex* structure exhibits loadings of the same item on different factors. However, such patterns are very difficult to detect with any rotation criterion conventionally applied in EFAs (an exception to some degree is the bi-factor criterion described by Jennrich & Bentler, 2011). Hence, if EFA was applied to a *factorial complex* structure the results may well be diverging. In *factorial complex* structures, the symptoms variances are expected to be caused by specific linear combinations of multiple factors. Such a principle suggests a dimensional structure instead of a categorical one. EFA analyses hardly ever retrieve an interpretable dimensional structure. Contrarily, NMDS models any data structure dimensionally. Since NMDS solutions can be looked at from a categorical viewpoint also, it represents an excellent tool to re-evaluate the previous exploratory factor analytic results. If NMDS results were evaluated in regards to a categorical structure (*factorial simple*), only groups of symptoms (i.e. distinct areas in the NMDS) would be interpreted. The location of within group items in their respective areas would be attributed to error and thus be considered random. In contrast, within group symptom locations in a dimensional (*factorial complex*) structure would be explicitly attributed to differences in the loading patterns on their respective factors (i.e. the differences in the linear combination of factors).

For our first analysis, pairwise correlations between the 21 items of the HAM-D_{21} were calculated for our sample of $N = 911$ depressive patients. Subsequently, the correlations were used as similarity measures for the NMDS analysis. The data for the second analysis originated from the pooled selection of studies included in two meta-analyses on the factor structure of the HAM-D_{17} by Bagby et al. (2004) and by Shafer (2006). However, six papers were not available to the authors and therefore were excluded: in five cases, the University of Zurich did not grant access to the required journals and in one case the book containing the study was out of print. Table 2 gives an overview on the 21 factor analytic results included in the NMDS analysis.

For this second analysis, we followed the guidelines of Loeber and Schmaling (1985), and Frick et al. (1993) who applied NMDS to conduct meta-analyses from factor analytic studies. These authors demonstrated that one does not necessarily have access to correlation matrices to create reliable similarity matrices. Each time two items revealed a loading greater or equal to 0.3 on the same factor, Loeber and Schmaling (1985) suggested to score the co-occurrence between those items with 1 else with 0. Hence, a symmetric co-occurrence matrix was created for each factor analysis that was included and each matrix consisted of the pairwise co-occurrence of the items. Thereafter, the matrices were summed up and each cell of the summed co-occurrence matrix was divided by the number of times both items were included in the same factor analyses. We applied two minor modifications to the procedure described by Loeber and Schmaling (1985). First, we included items only if their factor loadings were greater or equal to 0.4 (instead of 0.3). This first modification had to be applied due to missing reports on factor loadings smaller than 0.4 in some of the studies. Second, the elements of the co-occurrence matrix had not been divided by the denominator in the current manuscript, since 17 items were used in all of the previous studies[2].

[2] Two studies applied a longer version than the 17-item HAM-D: One study by Hamdi, Amin, and Abou-Saleh (1997) used a 21-item HAM-D version, whereas Reynolds, Kobak and Kobak (1995) applied a 23-item version of the HAM-D. Only the first 17 items were used in these cases.

Table 2

The results of factor analytic studies which were included in our second analysis of the HAM-D_{17} item structure

Author(s)	Sample	N. of Factors	*n*	year	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16	Item 17
Addington et al., time 1 (1996)	schizophrenic	7	112	1996	1	1,5	5	2	2	5,6	1,4	4	-3,6	7	1	7	1,7	-4,7	3	-6,1	6
Addington et al., time 2 (1996)	schizophrenic	7	89	1996	1,-3,-2,7	-3,7,-2	7	-3	-3	5,-3	-2	-2	-3,-2	1,-2	1	1	6	6	4	5,-4,1	1
Akdemir et al. (2001)	depressive	6	94	2001	1	2	2	3	3	3	5	5	1	1,2	2	1	6	4	4	1	6
Brown et al. (1995)	depressive	6	259	1995	3	3	3	1	5	5	6	3	1	1	1	2	6	5	4	2	4
Fleck et al. (1995)	depressive	6	79	1995	1	6	4,6	3	3	3	1	1	2	2	2	4	1	5	2	4	6
Gibbons et al. (1993)	depressive	5	370	1993	4,1	1	1	-5	-2	-2	1,4	4	1	1	1	-5	4	1	-	3,-5	-3
Hamdi et al. (1997)	depressive	6	100	1997	1	1	-3	3	-6	2	-4	1	2,6	3	4	5,4	5	-	4	2	6
Hamilton (1960)	depressive	4	70	1960	1	1	1	2	2	2	2	-3,1	3,2	3	3,4	2	4	1	-	2	1,2
Hammond (1998)	depressive	4	100	1998	4	-	-	2	2	2	3	4	1	1	1,4	3	3	-	3,1	3	-
Marcos & Salamero (1990)	geriatric	3	234	1990	2	-	2	3	3	3	2	2	-	2	1	-	1	-1	1	-	-
Michaux et al. (1969)	psychiatric	7	158	1969	2	2,5	5	1	1	1	2,7	7	2	2	4	3,6	6	6	4	3	-
O'Brien & Glaudin (1988)	depressive	6	183	1988	1	-6,1	1	-	4	4	2	-5	5	2,1	2	3	2	1	4	3	6
O'Brien & Glaudin (1988)	depressive	8	182	1988	3	-7	6	2	2	2	3	-1	1	3	6,5	4	5	7	6	4	8
Onega & Abraham (1997)	geriatric	4	206	1997	1	1	1	2	2	2	1	1	4	3	3	2	1	1	3	2	4
Pancheri et al. (2002)	depressive	4	186	2002	3	2	-	1	-	1	3	-	2	-	1	4,2	-	-	1	4	-
Ramos-Brieva & Cordero-Villafafila (1988)	depressive	5	115	1988	3	2,3	1,2	1	1	1	-4,3	3	2	2	2	-4	-4	-4	5	5	2
Reynolds & Kobak (1995)	mixed (psychiatric/ normal)	4	357	1995	1	1	1	2	2	2	1	1	4	4	4	3	1	1	4	3	4
Steinmeyer & Möller time 1 (1992)	depressive	6	223	1992	2	5	5	3	3	3	2	4	1	2	1	6	4	6	1	4	6
Steinmeyer & Möller time 1 (1992)	depressive	2	174	1992	1,2	2	2,1	1,2	1	1	1	-	2	1	1	1	1	-	-	-	-
von Giesen et al. (2001)	somatic, HIV	8	202	2001	7,1	6	1	2	2	2	3,8	7	6,8	1,6,7	1,5	8	8	5	4,6	4	3
Weckowicz et al. (1971)	depressive	4	52	1971	1	-3	-4	1	1	1	4	-2	2	2	2,3	1,4	3	1	3	1,4	3

Results

Results analysis 1

Figure 1 shows the two dimensional NMDS solution of all 21 HAM-D symptoms. The NMDS solution of our sample, which consisted of $N = 911$ patients diagnosed with depression, revealed a stress value of 0.225 in two dimensional space. The value was well below the critical stress value of 0.284 for unstructured data with comparable characteristics (Sturrock & Rocha, 2000).

At first glance, the HAM-D symptom structure did not look strictly categorical at all. Rather, the symptoms seemed to be ordered dimensionally in the two dimensional NMDS solution. By forcing a categorical structure, however, previous factorial findings could easily be identified within the structure. The often confirmed "insomnia factor", constituted by items 4, 5, and 6 was found in the upper right corner and the symptom-group was named "insomniac symptoms" accordingly. Adjacent to the "insomniac symptoms", the two "gastrointestinal" symptoms 12 (gastrointestinal) and 16 (weight loss) were found. Continuing counter clockwise, items 9 (agitation), 10 (psychic anxiety), and 11 (somatic anxiety) grouped together to an "increased activation" group of symptoms, followed by 13 (somatic symptoms general), 14 (genital symptoms), and 15 (hypochondriasis), which defined a more general "somatic symptoms" group.

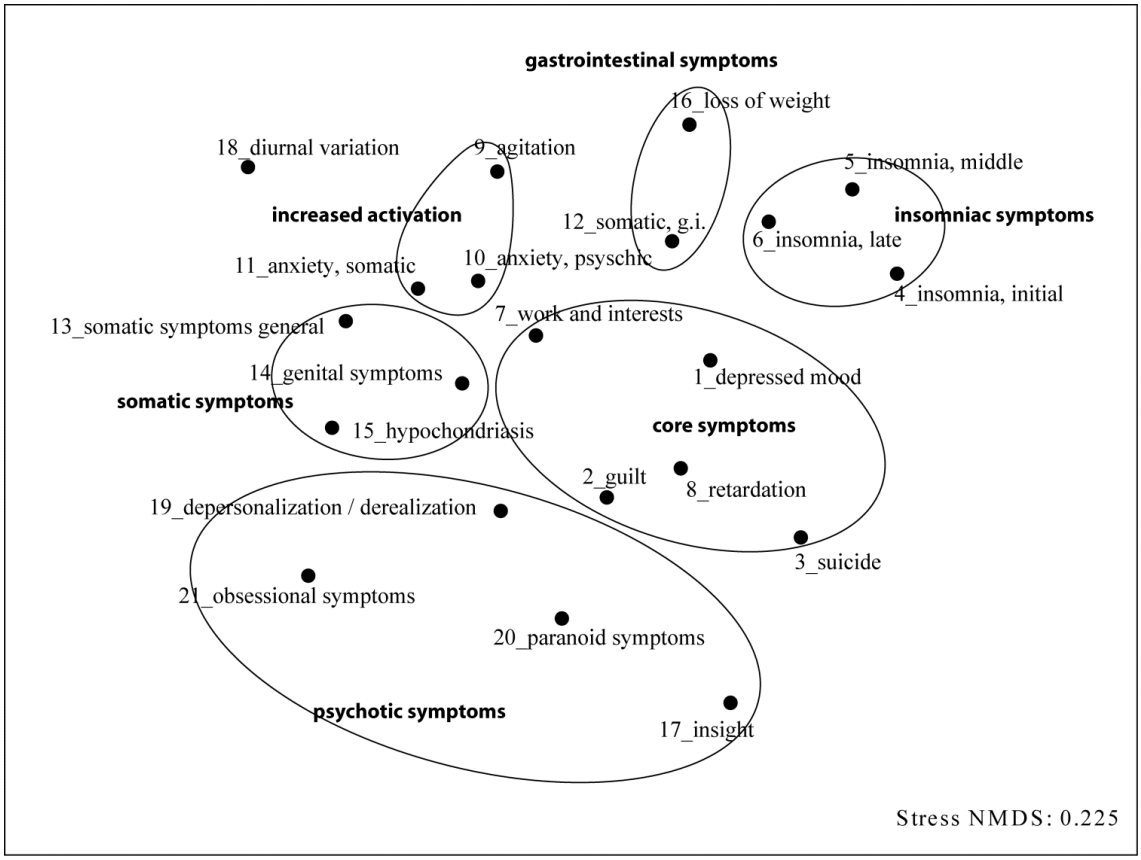

Figure 1. Two dimensional NMDS solution of the 21 HAM-D items based on the dataset consisting of depressive patients. Encircled items constitute a coherent group of symptoms.

Items 17 (insight), 19 (depersonalization/derealisation), 20 (paranoid symptoms), and 21 (obsessional symptoms) spanned a relatively broad area which was labelled as "psychotic symptoms". At last, items 1 (depressed mood), 2 (guilt), 3 (suicide), 7 (work and interests), and 8 (retardation) were all located in the centre. This last item group constituted a core factor of depression and consisted of motivational and cognitive symptoms mainly. So far, a strictly categorical interpretation was applied, since only groups of symptoms had been considered.

From a dimensional perspective, the locations of the clusters to each other as well as the individual locations of the symptoms are noteworthy. Firstly, on the level of factors, the most general ordering principle was found along the *y*-axis which separated psychotic, cognitive/motivational and somatic symptoms. No similar significant ordering principle could be identified on the second dimension; however, the *x*-axis was essential for unfolding the variability within the somatic and the psychotic symptoms. Starting from the upper right corner, the adjacent groups of insomniac and gastrointestinal symptoms could be categorized as psychovegetative symptoms; by expanding along the *x*-axis, psychovegetative symptoms, anxiety and general somatic symptoms combined to a broad somatic cluster, which laid on a quarter of a circle segment around the cognitive/motivational core symptoms. The circular arrangement of the items persisted for the psychotic symptoms. Their location on the lower segment of the circle maximized the distance to the psychovegetative symptoms.

Secondly, on the level of symptoms, the transition between somatic and psychotic symptoms was marked by item 15 (hypochondriasis). Specifically, hypochondriasis was located between item 13 (general somatic symptoms), and items 19 (depersonalization/derealisation) and 21 (obsessional symptoms), but also relatively close to items 10 and 11 (psychic/somatic anxiety). Furthermore, all the psychotic symptoms tended to be located closer to the core symptoms (cognitive/motivational) than to the somatic symptoms. Item 9 (guilt) was located in the transition between core symptoms and psychotic symptoms whilst item 7 (work and interest) was located between the core symptoms, anxiety, and general somatic symptoms. These findings are in good accordance with theoretical considerations which are given in the discussion in detail.

Located in the upper left corner, diurnal variation did not share much variance with any other HAM-D item. Its location was characterized by the absence of direct neighbours and by maximized distances to the core symptoms. Hence, it is safe to assume that diurnal variation yielded substantially different information than the other HAM-D items. As Hamilton (1960) himself already pointed out, symptom 18 (diurnal variation) is a poor predictor for severity of depression. Its outlying position in the upper left corner confirmed that theoretical notion on an empirical basis: the distances to any of the other symptoms were maximized by its location.

Results analysis 2

Figure 2 shows the NMDS solution of the first 17 HAM-D items. Here, the input data consisted of the co-occurrence matrix of previous factor analytic studies (Table 2). The four distinct groups of symptoms "insomniac symptoms", "gastrointestinal symptoms", "increased activation", and "core symptoms" could still be identified. Moreover, the dimensional ordering of the symptom-groups along the *x*-axis remained the same. The most pronounced differences certainly concerned the symptoms of the general somatic symptom-group (which were noticeably scattered) and item 17 (loss of insight), which changed its location to the upper left corner of the NMDS configuration, presumably because the other symptoms of the psychotic symptom-group were no longer present in the data. Also, the core symptoms were spread along the *y*-axis, and covered the area where the psychotic/delusional symptoms were positioned in the 21-items HAM-D (Figure 1).

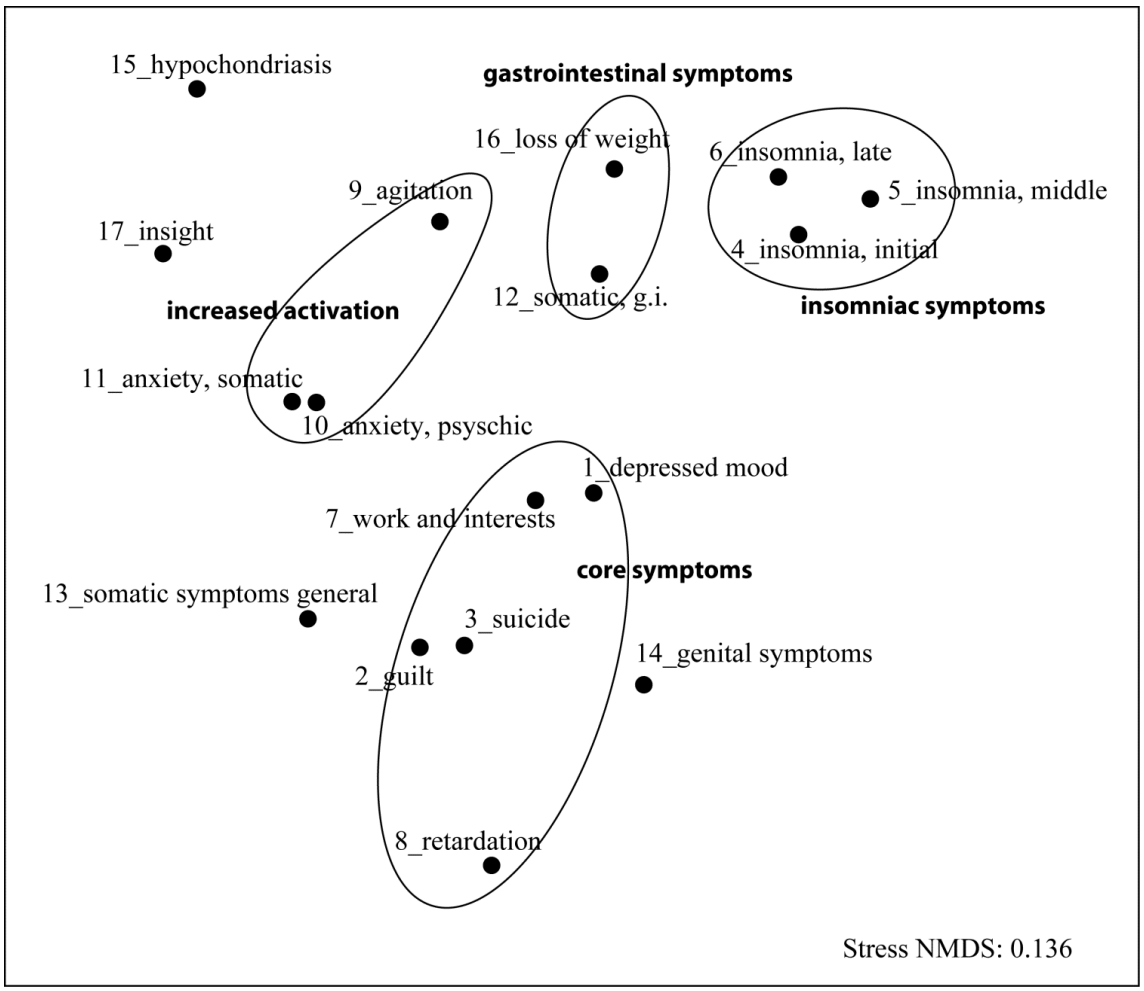

Figure 2. Two dimensional NMDS solution of the 17 HAM-D items based on the dataset of co-occurrences in factor analyses. Encircled items constitute a coherent group of symptoms.

Discussion

The discussion will be structured according to four different viewpoints we consider relevant: First, the structure of the HAM-D_{21} items will be discussed in detail and the main limitations of the analysis will be pointed out. Second, the solutions of the 17- and the 21-item version of the HAM-D will be compared. Third, the retrieved item structures will be compared to and integrated in the context of previous findings of HAM-D item structure and the broader theory of depression. Fourth, the practical implications of our findings will be pointed out. This fourth section contains a framework to analyse revised items in the context of the full HAM-D breadth and a proposal on how to depict the information from the HAM-D for clinical practice.

General discussion of HAM-D_{21} item structure

First of all, we would like to return to the individual locations of those HAM-D symptoms that were located in a transition region between symptom-groups (Figure 1). From a theoretical point of view, the transitional locations of the items guilt and hypochondriasis were probably guided by the same principle. On the one hand, both items require hallucinations or delusions to achieve the highest score. Hence, a location in the near of the psychotic symptoms is plausible. On the other hand, pathological guilt refers to cognitive distortions (thus presumably the close distance to cognitive symptoms) whereas a bodily self-absorption or a preoccupation with health (lower hypochondriasis scores) may refer to general health troubles (thus presumably the close distance of hypochondriasis to general somatic symptoms). Therefore, the actual locations of hypochondriasis and guilt in the NMDS solution (Figure 1) concur well with the theoretically derived ones, which would locate the items between the somatic and the psychotic factor, and the cognitive/motivational and the psychotic factor respectively.

The close range of item 7 (work and interest) to symptoms of increased activation and somatic symptoms is more difficult to explain. The HAM-D associates a considerably heterogeneous list of symptoms with item 7 (work and interests) what may be key to understanding its location in the NMDS solution. On the one hand, the patients' subjective states (loss of interest, feelings of incapacity or indecision), and on the other hand, the patients' activities (decreased productivity or social activities, inability to work) defines the score of item 7. However, an observed decrease in activity does not necessarily have to originate from loss of interest, feelings of incapacity or indecision. Just as well, it may relate to general health problems or severe anxiety. In this respect, work and interest lacks the innate specificity of the other HAM-D items. It probably is a good estimate for the global severity of illness though (work and interest is included in all of the HAM-D short scales (Table 1)) – probably because of its association with the patients' general level of functioning.

The centred position of the symptoms work and interest (item 7), depressed mood (item 1), retardation (item 8) and guilt (item 2) in the HAM-D_{21} NMDS solution (Figure 1) indicates

the items major contribution to the severity of depression: because the centre of a configuration is characterized by minimal distances to any other location, items in the centre of an NMDS solution generally reveal the highest correlations to all other items. These core items of depression were identified before and have been integrated in most of the Subsets of the HAM-D (Santen et al., 2008; Bech et al., 1975; Maier & Philipp, 1985; Gibbons et al., 1993; Evans et al., 2004). The item suicide, which is located in the same region, is included in the subscales of two of the five studies on subscales of HAM-D (Gibbons et al., 1993; Santen et al., 2008). Santen et al. (2008) included suicide because of its sensitivity to change in active treatment responders and Gibbons et al. (1993) because of its IRT characteristic to separate severely ill depressive patients. Hence, we included the item suicide in the core symptom-group, although its location is not as central as the locations of the other core symptoms. There is one item (anxiety, psychic), we rejected from the core symptom-group despite its inclusion in the subscales and despite its close distance to the core symptoms in the NMDS solution. Firstly, its close distance to the item anxiety, somatic and to agitation shapes a coherent cluster of increased activation, and secondly, although closely related to depression, anxiety is a core concept of another disorder.

As revealed by our NMDS solution, the covariance structure among HAM-D symptoms allows good semantic interpretation even for items not closely related to depressive core symptoms and despite low diagnostic diversity in our sample. Each item, except for diurnal variation which shall be excluded from the following statements, could be allocated to a compound of similar items and attributed to a theoretically meaningful symptom-group.

Generally, a circular cluster-structured NMDS solution, as it was obtained for the HAM-D, satisfies three principles: Firstly, each cluster must exhibit positive covariance with the construct being measured (otherwise it would “fall out” of the circle surrounding the core symptoms). Despite many studies propagating a multidimensional structure of the HAM-D, Cronbach’s alpha, as a measure of internal consistency, has repeatedly been found to exceed 0.70 (Bagby et al., 2004) indicating adequate reliability of the scale. Adequate internal reliability also supports the hypothesis of positive covariance with severity of depression and thus indicates a systematic principle in the circular ordering of the symptom groups. Secondly, each item must exhibit a pattern of covariances similar to other items of the same group because otherwise a narrow area of within group symptoms would not emerge. Thirdly, similarity between groups of symptoms must follow distinct principles to result in dimensionally ordered group positions. If the third principle was not met, the NMDS configuration could still look the same as Figure 1. In fact, the principle cannot be verified by the first NMDS result alone. However, it would be highly unlikely to replicate this very structure in a completely different set of data (moreover with different similarity measures); but this is exactly what we obtained from our second analysis with the co-occurrence matrix from previous factor analyses (Figure 2). Moreover, the circular ordering of the symptom groups around the core symptoms and the centrality of the core

symptoms are in good accordance with previous facet theoretical notions on the HAM-D symptom structure (Steinmeyer & Möller, 1992). By applying the wording of Steinmeyer and Möller (1992), there is a distinction of the groups with respect to a centrality facet: it divides the space in inner (core symptoms) and outer regions (other symptom-groups). Furthermore, Steinmeyer and Möller (1992) also described an "aspect" facet, which is concurrent with the circular ordering of the symptom-groups in the outer regions in our NMDS solution of the HAM-D (a more detailed comparison of our NMDS solution with the solution by Steinmeyer and Möller (1992) is given in a following section).

Obviously, different language versions of an inventory might lead to differences in the structure of the symptoms. Thus, even though these differences are usually small, one limitation of this study originates in the use of the German version of the HAM-D. However, the main limitations of our study are closely related to limitations inherent in NMDS analyses. The locations of items in NMDS solutions are not depending on the "true" underlying structure alone. They are depending on the dimensionality as well as on random error in the proximity matrix also. Hence, it is possible that some items might radically change their location in two independent solutions due to a good fit on two different locations (and a bad fit on any other location in between) – especially items on the edge of the structure. To address this issue, we applied bootstrap-methods for the HAM-D_{21} NMDS solution and identified two items (agitation / hypochondriasis) as items with two possible locations. For the item agitation, one location is on the edge of the increased activation cluster (Figure 1) the other would be within the psychotic symptom-group between paranoid symptoms and obsessional symptoms (not shown). For the item hypochondriasis, one location is in the transition between somatic and psychotic symptoms, the other would be within the psychotic symptom-group between obsessional and paranoid symptoms (not shown). The bootstrapped NMDS solution was favoured because bootstrapped NMDS solutions show more robust results (Bühler & Läge, 2013). Also, the bootstrap NMDS solution was predominant in regards to interpretability, which is considered an important quality criterion according to Borg and Staufenbiel (2007). A second limitation arises from the relational model inherent to NMDS: The configuration does not allow any statements about internal consistency of the HAM-D. Therefore, correlations between symptom groups cannot be estimated, not even for two opposing groups.

Comparison of the HAM-D_{21} and the HAM-D_{17} item structure

Although derived from completely different data, there is good accordance between the two NMDS solutions of the 21-items HAM-D from 911 depressive patients and the 17-items HAM-D from previous factor analyses. Four of five symptom clusters have been replicated; moreover, the dimensional ordering between symptom groups remained constant from one solution to the other. This result is especially astounding given the small sample size of the second analysis (data from 21 factor solutions) and by considering the number of distances between the 17 items

($n_{distances}$ = 136), the heterogeneity of samples used in those factor studies (in regards to culture and disorder) and the simplicity of the co-occurrence measure applied. The measure of co-occurrence described by Loeber and Schmaling (1985) is a simple measure because it does neither consider the level of factor loadings, nor the sample size of the underlying study or the number of significant factor loadings of the items. Hence, estimates of co-occurrence are prone to larger random errors. In the present study, it seems safe to assume that a good solution could be attained from the co-occurrence matrix though, given the high degree of accordance between the two NMDS solutions, which were derived from completely different data and similarity measures.

The few differences between the structures in Figure 1 and Figure 2 may be due to the omission of the psychotic/delusional symptoms. By omitting the last four items, a large part of variability in the data was omitted as well, which resulted in the loss of anchor points for the items guilt, hypochondriasis and insight. Following the same principle, the core symptoms in Figure 2 were no longer bound to the centre anymore. The core symptoms in the co-occurrence based NMDS were located on the lower end of the structure, yet still centred horizontally. The lack of the vertical centring may be found in the high covariance between the (mostly) somatic symptoms in the upper half of the structure (Figure 2). This somatic covariance might overshadow the central role of the core symptoms. Hence, considering only the first 17 items of the HAM-D, it may well be that a balancing counterweight to the somatic symptoms was missing – specifically to the insomniac and gastrointestinal symptom-groups. The psychotic symptoms, although rarely found, may therefore constitute a complementing group of symptoms.

The assumption that the psychotic symptoms act as a counterweight to the psychovegetative symptoms (insomniac and gastrointestinal symptoms) is supported by the results of Faries et al. (2000). Comparing the effects of fluoxetine vs. placebo and tricyclic antidepressants vs. placebo, the HAM-D_{21} scores outperformed the HAM-D_{17} scores in their large scale meta-analysis. Moreover, comparing the scores of the full scales to the HAM-D short scales (e.g. Maier & Philip, 1985; Bech et al., 1975; Gibbons et al., 1993), both full scales performed worse than the short scales, counter indicating a simple effect due to an increase in the number of items. It has to be noted though, that effect sizes between active treatment and placebo groups rely on change of symptomatology rather than on state reliability.

Hamilton (1960) recommended excluding items 18 – 21 in the final computation of the severity score for different reasons. He reported that the Items 19 – 21 occurred very infrequently and thus he saw no reason in including them in the total. Item 18 was excluded on different grounds. He did not believe that diurnal variation measured depression intensity, but rather that it defined the type of depression. Hamilton's notions on the different categories of those remaining 4 items are directly reflected in the NMDS solution of the HAM-D_{21}. Indeed, favouring a lonely spot in the upper left corner of the configuration, diurnal variation seems to

be measuring a different construct than any other item. This finding is in line with Hamilton's original assumption and with the opinion of many authors since, who pointed out the peculiar role of diurnal variation within the HAM-D.

While an interpretation of diurnal variation's location in the NMDS solution (Figure 1) in regards to its neighbours does not easily come to mind, items 19 – 21 are grouped together between insight, hypochondriasis, and guilt and thus shape a coherent cluster of psychotic symptoms. With improving distances from core symptoms, items 19 – 21 represent symptoms with increasing psychotic quality. This theoretically plausible location of the psychotic symptoms and their grouping together to a coherent symptom cluster indicates a valuable complement in the HAM-D questionnaire and should not be neglected.

The HAM-D item structure in the context of previous findings

Steinmeyer and Möller (1992) proposed an NMDS solution of the HAM-D_{17} items based on a sample of 223 patients with major depression. Although not strictly in accordance with our findings, many features of their NMDS solution were replicated. Specifically, they also identified the insomniac symptoms as a separate area, and those symptoms that we identified as core symptoms were positioned in close distance in the solution by Steinmeyer and Möller (1992) as well. However, some differences exist: The symptoms loss of weight and gastrointestinal symptoms did not define a coherent group in the NMDS solution retrieved by Steinmeyer and Möller (1992); a result that has clearly emerged in both our NMDS solutions. Also, the items somatic anxiety, somatic symptoms general, and genital symptoms are positioned much closer to each other in the Steinmeyer and Möller (1992) solution than it is the case in our co-occurrence based NMDS solution (Figure 1). However, in this regard, the solution of Steinmeyer and Möller (1992) is in close accordance to our NMDS solution of the HAM-D_{21} (Figure 1). The core symptoms of the HAM-D_{21} NMDS solution (Figure 1) are centred and thus concur well with Steinmeyer and Möller's (1992) notion of a centrality facet, however, they defined the set of core symptoms slightly different than we did. Moreover, the segmentation of the HAMD's symptoms according to an "aspect" facet, which is circularly ordered around the core symptoms, is concurrent with our results (Figure 1) as well. The encircled items in Figure 1, which can be grouped in "insomniac symptoms", "gastrointestinal symptoms", "increased activation", "somatic symptoms" and "psychotic symptoms", mirror the distinctions of different aspects of depression symptoms.

The major contribution of the NMDS analyses in Figure 1 and Figure 2 is that they might explain the diversity in factor analytic results over the decades. Both structures discard a *factorial simple* solution. As was discussed in more detail in the methods section, a *factorial simple* structure would have been indicated if distinct clusters of items were obtained. In contrast, in the current findings (Figure 1 & 2), a purely data driven clustering of items would yield ambiguous results, especially if different numbers of clusters were applied. Therefore, diverging

factorial simple structures have similar probabilities to emerge in EFAs and are prone to alter with only a slightly different set of data.

In confirmatory analysis a *factorial complex* structure could easily be integrated and tested. A factorial complex structure seems especially promising since several items in the HAM-D are likely to measure different concepts (e.g. Bagby et al., 2004; Zimmerman et al., 2005). However, the only confirmatory factor analysis of the HAM-D known to the authors is a study by Cole et al. (2004), which implements a *factorial simple* structure. Cole et al. (2004) concluded that the model did not reach acceptable fit indices; one likely explanation could be that their factorial simple model was not able to reproduce the dimensional structure of the data.

Bagby et al. (2004) noted that the items guilt and hypochondriasis violate basic measurement principles, because they unite two different concepts. In the item guilt, self-reproach and ideas of guilt mark the less severe categories and stay closely in line with cognitive distortions, whereas delusions and hallucinations of guilt, which mark the more severe categories, rather can be seen as psychotic features. As Bagby et al. (2004) noted, in the latter categories, guilt may not be represented with higher magnitude but a second, more severe concept (psychotic feature) may determine the item score. Given the position of the item guilt in the NMDS solution, however, it rather indicates a well suited item to measure severity of depression, a finding that is supported both, by the items usage in HAM-D short versions (Maier & Philipp, 1985; Gibbons et al., 1993; Bech et al., 1981) and by findings from IRT-analyses (Santor & Coyne, 2001; Evans et al., 2004).

To conclude that psychotic features generally indicate increased severity in depression would be misleading though. As can be seen in the NMDS solution (Figure 1), strictly psychotic items are positioned farther apart from the centre, suggesting a diminished influence on depression severity. If not the psychotic characteristic of the item guilt, what is it that makes this item a good estimator for depression severity? Maybe it's not an increase in the *magnitude* of guilty feelings, but rather in the *persistence* of guilty feelings that relates closely to severity of depression. It seems reasonable to assume that on the continuum between slightly over-accentuated (self-reproach) and heavily distorted cognitions (hallucinations of guilt) the persistence of guilty feelings is increasing. Hence, the question asked by Bagby et al. (2004) whether a patient with hallucinations of guilt is feeling more guilt than a patient with simple ideas of guilt might focus the wrong concept. If actually persistence of guilty feelings was the main influence on severity of depression, the item guilt could cover the range quite reasonably.

Practical implications

In the context of research, the NMDS solution of the HAM-D proposes a framework, in which newly designed items could be interpreted in regards to the test as a whole – assuming the original HAM-D and the newly designed items both would be collected in the same sample. On the

one hand, if a strict unidimensional measure was pursued, the new items should be located within the cluster of the core symptoms. On the other hand, if a measure was sought with a similar breadth to the original HAM-D, the newly designed items should cover approximately the same area in NMDS space as the HAM-D items, but they should be more strictly separated. In this case, a configuration with distinct clusters should be pursued.

In practice, HAM-D scores are widely collected and accepted not only to document depressive states but also to assess the results of treatment (e.g. Helmreich et al., 2012). Despite the many pitfalls in psychometric properties of the HAM-D that recent publications have revealed, chances are, the HAM-D will endure. In this case, it may be worthwhile to develop a better way to structure the HAM-D for practical use. Especially, the full breadth of the symptoms and symptom-groups captured by the HAM-D should be made available. In this regard, a categorical approach becomes limited quickly as the diverging results of factor analytic studies have demonstrated. Contrarily, a profile score (the scores of each symptom separately) of the HAM-D may yield this full breadth of information. However, the important information about the covariance between the HAM-D symptoms could not be displayed in such a profile score and the partial information may be difficult to integrate to one overall assessment.

The herein proposed NMDS solution not only is a fruitful tool to explain the diverging factor analytic results, moreover, it could yield the information clinicians are longing for: Specifically, HAM-D items could be coloured directly within the NMDS configuration, according to the assessed severity values of individual patients (for example yellow = 1, orange = 2, red = 3/4). Such a coloured, individual symptom profile would not only contain information about the severity of depression in different aspects of depression (accumulation of coloured items in specific areas), but would also transmit information about the structure of depression (the position of items and symptom groups to each other). As Zimmerman et al. (2005) pointed out, one of the major advantages of the HAM-D is its capability to assess additional features associated with depression. They are frequently used by clinicians to select antidepressants (Zimmerman et al., 2004). Depicting item severity directly in NMDS solutions would easily grant access to this level of detail, moreover, it could provide beneficial information on the structure of the disease that cannot be accessed until yet.

Acknowledgements

This study was supported by Grant 12674.1PFLS-LS from the CTI (Federal's Comission for Technology and Innovation).

Better bootstrap NMDS analyses – confidence regions and improved location estimates in Nonmetric Multidimensional Scaling

Joël Bühler & Damian Läge

Department of Psychology, University of Zurich, Zurich, Switzerland

Submission status:

Submission pending.

Authors' contributions:

Joël Bühler: Development of the research question, review of the literature, development of the study design, execution of the analyses, interpretation of the results, writing of the manuscript

Damian Läge: Supervision and discussion of Joël Bühlers contributions, revision of the manuscript

Abstract

Nonmetric Multidimensional Scaling (NMDS) has proven to be a valuable tool to assess the structure of multidimensional data – even outside its original domain of psychophysics. However, its results have been noted with caution because of the imponderability of sample effects and methodological issues like local minima. To address these issues, confidence regions based on bootstrap methods were considered and evaluated. Additionally, the bootstrap distributions were used to derive improved location estimates. It is demonstrated that the bootstrap indeed is a valid method to compute confidence intervals in NMDS, as long as the percentile method is applied. The results from the improved location estimates are promising, and it is argued that they can be attributed to the reduction of systematic biases from proximity estimation and from local minima.

Keywords: Nonmetric Multidimensional Scaling, bootstrap, confidence regions

Introduction

Different variants of Nonmetric Multidimensional Scaling (NMDS) have been in use for over a half century and were applied in many fields of psychology. From NMDS' origins in the early sixties (e.g. Kruskal, 1964a; Shepard, 1962), where the method was primarily applied in psychophysics, it became a popular statistic in the late seventies with a wide variety of minimization algorithms such as iterative majorization (De Leeuw, 1977; Mathar & Groenen, 1991), ML estimation (Ramsay, 1977), and steepest descent (Kruskal, 1964b). Furthermore, the nonmetric transformation varied with the algorithms about isotonic regression (Commandeur & Heiser, 1993), rank-images (Lingoes & Roskam, 1973) and spline transformation (Ramsay, 1982), to name a few without claim of exhaustiveness. Additionally, some algorithms also diverged in their weighting functions which were used to scale the dissimilarities either for reasons of robustness (Läge, Daub, Bosia, Jäger, & Ryf, 2005) or for adjusting for individual differences (Carroll & Chang, 1970; Ramsay, 1982). Along with classical (Torgerson) scaling (Torgerson, 1952) and metric multidimensional scaling, it has been applied in many research areas within the field of psychology such as marketing research (e.g. Carroll & Green, 1997), clinical psychology (e.g. Läge, Egli, Riedel, & Möller, 2012), developmental psychology (Loeber & Schmaling, 1985) or neuropsychology (Abdi, Dunlop, & Williams, 2009). Moreover, multidimensional scaling has been applied in various research areas such as cross-cultural research (Padilla, Benítez, Sirec, & Flores-Galaz, 2012), the human sciences (Vanpoucke, Boermans, & Frijns, 2012), finance (Cox, 2012) and engineering (Qin, Wan, Duan, 2012) documenting its unbowed significance among established methods of analysis.

Despite its widespread use, there have always been concerns about the stability and generalizability of NMDS results. In this regard, the issue of outlier handling (Läge et al., 2005; Spence & Lewandowsky, 1989) and, especially in lower dimensionality, of local minima (Groenen & Heiser, 1996) are mentioned. Despite these reservations, the family of NMDS algorithms have proven to be a valuable tool for interpreting the structures of datasets. Substantial insight can be obtained, especially from NMDS solutions in two and three dimensional space, in which a visual representation of the results can easily be achieved.

The primary field of application of NMDS analyses has been in exploratory analyses. There have been suggestions for inference based tests to test hypotheses within and between models (a good overview is given by Borg & Groenen, 2005), but they have been sparsely used. The maximum likelihood algorithms of NMDS are generally well suited to test hypotheses and obtain standard errors of statistics. However, there are specific stumbling blocks (unidentified parameters and nuisance parameters; for details refer to Ramsay, 1982) that are challenging the estimation through asymptotic properties. Furthermore, ML MDS models assume independence of, and normal or lognormal distributed errors, which is too rigid for many datasets.

The bootstrap, a method for resampling described by Efron (1979), has proven useful in calculating confidence regions under minimal assumptions, and it has already been applied in classical multidimensional scaling (Abdi et al. 2009). Furthermore, the bootstrap has been suggested to calculate confidence regions in an NMDS framework by Heiser & Meulman (1983) and Weinberg, Carroll & Cohen (1984). In a series of analyses, Weinberg et al. (1984) compared bootstrap and jackknife confidence regions for their INDSCAL algorithm (Carroll & Chang, 1970) with confidence regions obtained from ML estimation in the MULTISCALE program (Ramsay, 1977). Additionally, they tested the adequacy of the resampled confidence regions with a Monte Carlo analysis. Regarding the shape and volume of the confidence regions, good accordance was found between the confidence regions based on resampling and the results of their Monte Carlo analysis. However, there were some limitations to the study by Weinberg et al. (1984). The authors neglected the issue of local minima in NMDS solutions, by setting the starting configuration to the known, true configuration in their Monte Carlo analysis. Additionally, their bootstrap sample was only of small size ($K = 21$), which contributed to the limitations regarding a general statement about the application of the bootstrap in NMDS.

The bootstrap was discussed in both studies (Weinberg et al., 1984; Heiser & Meulman, 1983) with respect to directly collected, pairwise similarity data (direct proximities; i.e. the subjects judged each pair of objects separately with respect to the objects' similarity). However, many applications of NMDS apply indirect similarity coefficients which are inferred from two-way two-mode data matrices (indirect proximities; e.g. Bühler, Keller & Läge, 2012; Egli, Riedel, Möller, Strauss, & Läge, 2009) and thus bring forth some specific considerations concerning the data. Firstly, independence among the similarity estimates can no longer be assumed. Secondly, individual structures cannot be computed (as is done in INDSCAL for example (Carroll & Chang, 1970)) because the aggregated values of the two-way two-mode data result in one single proximity data matrix only. These considerations, along with the limitations in the study by Weinberg et al. (1984) demand a systematic evaluation of the bootstrap in NMDS.

The current study was split in two separate analyses, each of which focused on one specific aim. Firstly, a systematic evaluation of the bootstrap procedure in an NMDS framework was pursued. The bootstrap enables for the computation of confidence regions, which are essential, if one of the major critiques about NMDS should be addressed: interpretations of NMDS solutions, i.e. the grouping of objects or the individual locations of objects, can easily be criticized for unknown sample effects and methodological difficulties. Abdi et al. (2009) gave an illustrative example for the usefulness of MDS confidence regions in the area of brain image data analyses to address these issues. Confidence regions in NMDS analysis are also especially promising in the area of psychopathology. The symptom structure of different mental disorders is of great concern to many researchers in psychology and psychiatry. However, conventional exploratory methods like exploratory factor analyses cannot account for dimensional symptom

structures (factorial complex structures), whereas NMDS can (Bühler et al., 2012); the lack of uncertainty estimators has hindered a widespread application in this area to date though.

The second aim of our study was to evaluate an extended procedure to NMDS analyses which we hypothesized would improve the fit of the results by reducing two different sources of error. The first source of error is a systematic error of aggregated proximities (indirect proximities) which has not been considered yet in NMDS analysis of two-way two-mode data to the best of the authors' knowledge. This proximity bias is described in detail in the conclusion of analysis 1. The second source of error is an unsystematic error, which is introduced by local minima. The extended procedure is described in detail in the introductory section of analysis 2.

The software ProDax (Oberholzer, Egloff, Ryf, & Läge, 2008) was used to conduct all NMDS analyses in the current study. As a default, ProDax applies the robust algorithm RobuScal to calculate NMDS solutions (Läge, Daub, Bosia, Jäger & Ryf, 2005). RobuScal relies on a weighted loss function to calculate NMDS solutions. In this loss function, the weights decrease with increasing error between disparity and distance. However, since generalizability of the results to other NMDS algorithms such as PROXSCAL (Commandeur & Heiser, 1993; Meulman, Heiser, & SPSS, 1999) or ALSCAL (Takane, Young, & De Leeuw, 1977) was desired, the weighting function of RobuScal was disabled in the current study.

Methods

Both analyses were based on the same set of simulated data. However, since the topics of our two aims vary greatly, we chose to report our results in two separately labelled sections.

Simulation procedure

The Monte Carlo analysis was based on the "European city distances" example used by Borg & Groenen (2005) in which the distance matrix of ten European cities was analysed. However, preceding the actual simulation procedure, a 2 dimensional NMDS solution of the distances was computed to reduce the actual (3 dimensional) distances to a 2 dimensional mapping. Subsequently, the object vectors for the first dimension ($\boldsymbol{y_1}$) and for the second dimension ($\boldsymbol{y_2}$) of the resulting configuration ($\boldsymbol{\theta}$) were extracted. These vectors were then used to create the data for our Monte Carlo Analysis. This procedure had two advantages. Firstly, it ensured that the true configuration was indeed a metric structure in two dimensional Euclidian space (and not a structure on the surface of an ellipsoid) and, secondly, that a two-way two-mode data table could be constructed easily, as is explained in detail below.

It is obvious that $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$ transform back into the distance matrix $\boldsymbol{D}(\boldsymbol{\theta})$ if a two-way two-mode data set only consisted of these two vectors and if the dissimilarities were calculated as Euclidean distances. Also, it has to be noted that $\boldsymbol{\theta}$ represents only one configuration of an

infinite number of different configurations which all solve the given NMDS problem in two dimensional space. Because NMDS solutions are invariant to translation, scaling, rotation and reflection, each configuration $\boldsymbol{\theta}_i$ that satisfies

$$\boldsymbol{\theta}_i = a \cdot \mathbf{1} + b \cdot \boldsymbol{\theta}\boldsymbol{Q}, \tag{1}$$

where $\mathbf{1}$ is an $n \times k$ dimensional matrix of ones, $\boldsymbol{Q}$ is any orthogonal matrix and a and b are real valued numbers, is a solution for the given NMDS problem.

To reduce the complexity of comparing $\widehat{\boldsymbol{\theta}}$-values with a given $\boldsymbol{\theta}$, the distance matrix $\boldsymbol{D}(\boldsymbol{\theta})$ was computed. Because Euclidean distances are varying under translation, rotation and reflection the above stated dependency of $\boldsymbol{\theta_i}$ on $\boldsymbol{Q}$, a and b reduces to

$$\boldsymbol{D}(\boldsymbol{\theta_i}) = b \cdot \boldsymbol{D}(\boldsymbol{\theta}) \tag{2}$$

Thus, to compare $\widehat{\boldsymbol{\theta}}$-values with a given $\boldsymbol{\theta}$, one needs only account for b, which, obviously can be accomplished by any normalizing transformation, for example by dividing $\boldsymbol{D}(\boldsymbol{Y})$ by the Sum of its squared elements.

As described in the introductory section, our primary interest in this study focused on confidence regions in NMDS based on two-way two-mode data (indirect proximities). Therefore, a two-way two-mode data table needed to be constructed which yielded the same NMDS solution as $\boldsymbol{D}(\boldsymbol{\theta})$. Of course, any two-way two-mode data set, which replicates the vectors $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$, reveals an NMDS solution that satisfies equation (1). Hence the dataset

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{y}_1^1 & \boldsymbol{y}_2^1 & \dots & \boldsymbol{y}_1^{N/2} & \boldsymbol{y}_2^{N/2} \end{bmatrix}$$

will lead to

$$\boldsymbol{D}(\boldsymbol{S}) = \left(\frac{N}{2}\right)^{1/2} \cdot \boldsymbol{D}(\boldsymbol{\theta}) \tag{3}$$

which satisfies equation (2) and thus will yield the same NMDS solution as $\boldsymbol{D}(\boldsymbol{\theta})$, if the metric of the proximities are Euclidean distances.

The vectors $\boldsymbol{y}^i$ from $\boldsymbol{S}$ could for example be conceived as estimates for the longitudinal ($\boldsymbol{y}_1^i$) and the latitudinal ($\boldsymbol{y}_2^j$) position of the cities. Assume that each vector represents the location estimates from one person and that two groups of persons existed: One group estimated the distances with respect to longitude and the other group with respect to latitude. Such data structures are very common in psychology data and they reveal systematic variance on three different levels of the data. Firstly, there is a level of systematic between group variance (the differences in the true values of longitudinal and latitudinal distances). Secondly, there is random variance

due to the sampling from the population (assume we don't know according to which criteria a person will estimate the distance between the cities beforehand). And thirdly, there is random within group variance, namely the error on the distance estimates given by each person.

To reflect the two random sources of error, the simulation of our data was a two-step procedure: In the first step, the number of persons in group one (the persons with the longitudinal distance criterion) was randomly chosen from a binomial distribution

$$n_{y_1} = \mathrm{B}(N, 0.5)$$

and the number of persons in group two (the persons with the latitudinal distance criterion) was then inferred as

$$n_{y_2} = N - n_{y_1}$$

where N denotes the total number of simulated persons. Since a binomial distribution with parameter 0.5 was chosen, the expected value of the number of estimators for both n_{y_1} and n_{y_2} were $N/2$, which assured unbiased solutions with regard to $\boldsymbol{\theta}$.

In the second step the "estimates" for each person were drawn from multivariate normal distributions separately for the two groups to simulate the within group error on the estimates.

$$\widehat{\boldsymbol{y}}_i \sim \mathrm{MN}(\boldsymbol{y}_1, \sigma^2 \cdot \boldsymbol{I}), \quad i = 1 \dots n_{y_1}$$

and

$$\widehat{\boldsymbol{y}}_j \sim \mathrm{MN}(\boldsymbol{y}_2, \sigma^2 \cdot \boldsymbol{I}), \quad j = 1 \dots n_{y_2}$$

Here $\boldsymbol{I}$ denotes the identity matrix and σ^2 the random error variance on the estimates.

To increase generalizability, the sample size was systematically varied from $N = 50$ to $N = 800$ by doubling the size at each following condition, and the standard deviation was varied in three steps from $\sigma = 0.2$ via $\sigma = 0.5$ to $\sigma = 1$ resulting in a total of fifteen (3 x 5) experimental conditions. For each condition, 100 iid sets were simulated ($\boldsymbol{S}_1, \dots, \boldsymbol{S}_{100}$). These sets constituted the backbone of our Monte Carlo analysis in which the confidence intervals from the bootstraps were evaluated. Table 1 shows the summed rank deviations between the distances' rank-orders of $\boldsymbol{D}(\boldsymbol{S})$ and $\boldsymbol{D}(\boldsymbol{\theta})$ across the different experimental conditions and averaged over 100 sets in each cell to illustrate the effect of the experimental condition on the distance matrix.

To calculate the confidence intervals, 1000 resamples of size N were drawn with replacement from each two-way two-mode dataset to estimate the bootstrap distributions ($\boldsymbol{S}^*_{l_1} \dots \boldsymbol{S}^*_{l_{1000}}$). Please note that NMDS had not yet been applied to the data. Finally, NMDS was applied to all constructed sets (to the top level sets $\boldsymbol{S}_l$ and to the resampled sets $\boldsymbol{S}^*_{l_i}$) and the distance matrices of the resulting configurations were computed. Thus, the distance matrix $\boldsymbol{D}(\widehat{\boldsymbol{\theta}}_l)$ and the bootstrap cdfs ($\hat{G}_{ij}$), which were given by the distributions of the individual distances in the bootstrapped NMDS solutions $d(\widehat{\boldsymbol{\theta}}^*_l)_{ij}$, could be obtained for each top level set $\boldsymbol{S}_l$.

Table 1

Mean rank deviations between D(S) and D(Y) across experimental conditions

N	$\sigma = 0.2$	$\sigma = 0.5$	$\sigma = 1$
50	63.44	115.04	242.86
100	46.46	85.56	180.24
200	35.28	60.64	130.34
400	25.18	44.08	95.08
800	21.94	36.12	73.68

Analysis 1

General procedures Analysis 1

In the first analysis, the general appropriateness of the bootstrap procedure in NMDS was evaluated. The analysis was split in two successive sections (a. and b.) because two different procedures to estimate confidence regions in NMDS were assessed. In analysis 1a, confidence intervals for each distance in the NMDS solution $\boldsymbol{D}(\widehat{\boldsymbol{\theta}})$ were calculated, where $\widehat{\boldsymbol{\theta}}$ denotes the NMDS solution of one simulated dataset ($\boldsymbol{S}_l$). Since the distances matrix is defined as a symmetric matrix with diagonal elements $d_{ii} = 0$ and $d_{ij} > 0$, the confidence intervals for each distance satisfies $d(\boldsymbol{\theta})_{ij} \in \left[\hat{G}_{ij}^{-1}(\alpha),\ \hat{G}_{ij}^{-1}(1-\alpha)\right]$ and $d(\boldsymbol{\theta})_{ii} = 0$. The distribution $\hat{G}_{ij}^{-1}$ is estimated from the bootstrap distribution $d\left(\widehat{\boldsymbol{\theta}}^*\right)_{ij}$ and α simply denotes the $100 \cdot \alpha$ percentile of the distribution. In analysis 1b, the confidence regions were obtained directly from the bootstrap distributions of any objects' location in *k*-dimensional space. Hence, each objects confidence region is defined as a *k*-dimensional region.

Both approaches reveal strengths and weaknesses when it comes to calculating confidence regions from bootstraps in NMDS. First, consider the case of confidence regions on the basis of the objects' distances $d_{ij}(\widehat{\boldsymbol{\theta}})$, the procedure that was used in analysis 1a. If the confidence regions were considered on the level of distances, their shapes could be defined by the actual probability distributions of the distances and need not necessarily be defined in the lower NMDS space. Such a model seems appropriate if a perfect fit cannot be achieved (which is hardly ever the case with empirical data) because in that condition scale invariance should not be assumed between the different NMDS spaces.

However, one possible flaw arises from constraints in Euclidian distance metrics. It is the metrics' property of symmetry and non-negativity ($d_{ij} = d_{ji}, d_{ij} > 0$) that needs special attention. Figuratively speaking, one never knows, which of two objects is "left" in an NMDS solution and which is "right". Euclidean distances simply do not reflect such a property. Of course, if the structure as a whole was transformed this would be irrelevant: the solution is invariant towards rotation, translation and scaling. But if multiple NMDS solutions projected into

the same space were considered, it would matter greatly if two objects swapped their position, since this would necessarily influence the distances between other objects as well. Thus, if the bootstrap distributions were calculated independently for each distance, the cases in which the objects swapped their positions could not be separated from those where the objects did not swap their positions.

Second, consider confidence regions directly computed from object locations in NMDS solutions, the procedure that was used in analysis 1b. There is one obvious disadvantage: the solutions inevitably need to be standardized, since NMDS solutions are invariant towards rotation, translation and scaling. One approach to standardization is the projection of the configurations into a standard *k*-dimensional space. The so called procrustean transformation can be achieved in multiple ways, as suggests the literature on the topic (Gower, & Dijksterhuis, 2004). It has been suggested that the resulting confidence region follows some elliptic function (e.g. Heiser & Meulman, 1983; Weinberg et al., 1984). However, elliptic confidence regions cannot fully represent the influence of the loss function on these confidence regions as soon as an NMDS solution cannot be fit perfectly. Moreover, any procrustean transformation must assume an explicit error model (for example that the error in location estimates is normally distributed), which may not be adequate in some cases.
Then again, there are positive effects worth mentioning, if the confidence regions are computed directly in the dimensionality of the NMDS solution. Due to a common space to all configurations, one does not have to worry about symmetry or non-negativity – within that reference space, any segment between two points has both length and direction.

As a conclusion, confidence regions calculated directly in a reference space discard the flaws of confidence intervals based on distance matrices and vice versa. Therefore the bootstrap was applied to NMDS analyses in two different contexts. In the first context, confidence regions were calculated based on distances ($d_{ij}(\widehat{\boldsymbol{\theta}}^*)$; distances context, analysis 1a), while in the second context, confidence regions were calculated based on configurations ($\widehat{\boldsymbol{\theta}}_i^*$; configuration context, analysis 1b).

Procedures Analysis 1a

In analysis 1a, a confidence interval was computed for each distance in each set. Since there were 100 sets per experimental condition and 45 distances in each set, a total of 4’500 confidence intervals resulted per experimental condition. The confidence interval was set to a $1 - 2\alpha$ Level of 95%. Three different methods were used to compute confidence regions: the percentile method, the bias corrected method (BC) and the bias corrected and accelerated method (BCa) (Efron & Tibshirani; 1986).

Table 2

Percentage of confidence intervals that included the true distance value $\boldsymbol{d}_{ij}(\boldsymbol{\theta})$

N	$\sigma = 0.2$			$\sigma = 0.5$			$\sigma = 1$		
	Pc	BC	Bca	Pc	BC	Bca	Pc	BC	Bca
50	93.1%	89.1%	88.8%	95.4%	89.6%	89.2%	92.3%	83.5%	84.0%
100	92.5%	88.3%	88.2%	95.3%	89.9%	90.1%	94.5%	87.0%	87.1%
200	90.5%	84.2%	84.4%	95.7%	89.7%	89.4%	95.8%	89.9%	90.0%
400	90.8%	83.6%	84.1%	96.7%	88.9%	89.1%	97.6%	90.0%	90.3%
800	91.7%	79.9%	81.1%	96.5%	83.9%	84.6%	97.5%	89.5%	89.7%

Note. The results were structured by experimental condition and calculation method. The three calculation methods are structured column wise: the percentile method (Pc), the bias corrected method (BC), and the bias corrected and accelerated method (BCa).

Results Analysis 1a

Table 2 shows the percentage of bootstrapped confidence intervals that included the true distance value $d_{ij}(\boldsymbol{\theta})$. Astonishingly, the percentile method outperformed the more general BC and BCa methods in all experimental conditions: the percentages of confidence intervals that contained the true distance value $d_{ij}(\boldsymbol{\theta})$ were closer to the defined $1 - 2\alpha$ Level of 95% for the simple percentile method in every experimental condition.

Table 3 shows the effect of calculation method on the mean spread between upper and lower bounds of the confidence intervals. There is no systematic difference for the three calculation methods. Hence, the superiority of the percentile method in computing confidence intervals for NMDS distances can't be attributed to a mere effect of larger confidence intervals.

Table 3

Mean spreads between upper and lower bounds of the confidence intervals structured by experimental condition and calculation method

N	$\sigma = 0.2$			$\sigma = 0.5$			$\sigma = 1$		
	Pc	BC	Bca	Pc	BC	Bca	Pc	BC	Bca
50	0.356	0.357	0.360	0.674	0.672	0.677	1.201	1.158	1.175
100	0.250	0.251	0.253	0.491	0.489	0.492	0.959	0.936	0.948
200	0.179	0.178	0.179	0.355	0.355	0.357	0.737	0.731	0.738
400	0.127	0.127	0.128	0.255	0.253	0.255	0.558	0.555	0.559
800	0.099	0.096	0.097	0.195	0.190	0.192	0.415	0.411	0.415

Note. The three calculation methods are structured column wise: the percentile method (Pc), the bias corrected method (BC), and the bias corrected and accelerated method (BCa).

Procedures analysis 1b

In experiment 1b, the confidence regions were calculated based on standardized configurations. For this purpose, we integrated a suggestion by Abdi et al. (2009), who describe the projection of individual matrices on a compromise map in a classical MDS framework. We exploited the property that $\boldsymbol{D}(\boldsymbol{\theta})$ yields a classical multidimensional scaling solution without error, since it already represents a distance matrix in two-dimensional space. The same property holds for any other distance matrix computed from any NMDS solution. Hence, $\boldsymbol{D}(\widehat{\boldsymbol{\theta}})$ can be projected on the same space as $\boldsymbol{D}(\boldsymbol{\theta})$ with the following projection matrix given by Abdi et al. (2009)

$$\boldsymbol{P}_{\theta} = \boldsymbol{V}_{\theta}\boldsymbol{\Lambda}_{\theta}^{-\frac{1}{2}} \tag{4}$$

where V and Λ denote the eigenvector and eigenvalues of the matrix $\boldsymbol{D}(\boldsymbol{\theta})$ respectively. The projection is then computed as the cross product

$$\widehat{\boldsymbol{\theta}}_{Z} = \boldsymbol{D}\left(\widehat{\boldsymbol{\theta}}\right) \times \boldsymbol{P}_{\theta} \tag{5}$$

The coordinates were treated as independent from each other to estimate the confidence regions about the true positions of our ten cities in the data set. Hence, for each city in each set, two confidence intervals were computed, one for its y_1- and one for its y_2-component. This resulted in a set of 20 confidence intervals per top level set and 2'000 confidence intervals per experimental condition.

The bootstrapped distance matrices were projected in the same way as the distance matrices of the sets (Eq. 4 & 5). However, instead of $\boldsymbol{P}_{\theta}$, $\boldsymbol{P}_{\hat{\theta}}$ was used which was based on the eigenvectors and eigenvalues of the matrix $\boldsymbol{D}(\widehat{\boldsymbol{\theta}})$. The cdfs, from which the confidence intervals were computed, were given by the bootstrap distributions of the objects' y_1- and y_2-components.

Results analysis 1b

Table 4 yields the results of analysis 1b, in which the confidence intervals based on standardized configurations were evaluated. Please note that the percentage of correctly included true values is expected to be somewhat higher than the alpha level applied because we treated the components of the configuration as independent from each other. This implies a confidence region that is rectangular in shape (instead of an ellipse which is chosen for example in ML NMDS) and restricted to expand along the basis of the standardized space. For the confidence intervals based on standardized configurations, the same pattern of dominance of the percentile method over the BC and the BCa method emerges.

Table 4

Percentage of bootstrapped confidence intervals that included the true objects' configuration component $\boldsymbol{y_{ki}}$

N	$\sigma = 0.2$			$\sigma = 0.5$			$\sigma = 1$		
	Pc	BC	Bca	Pc	BC	Bca	Pc	BC	Bca
50	93.4%	89.1%	88.6%	96.1%	91.3%	91.1%	94.1%	84.1%	84.7%
100	92.7%	87.8%	88.0%	96.5%	91.7%	92.2%	95.2%	88.7%	88.9%
200	89.4%	83.9%	83.8%	96.1%	90.5%	90.5%	97.1%	90.7%	90.8%
400	91.7%	85.9%	86.4%	97.4%	89.0%	89.0%	97.9%	91.1%	90.9%
800	93.3%	84.7%	84.9%	97.0%	87.0%	87.2%	97.4%	89.3%	89.5%

Note. The true objects' configuration component $\boldsymbol{y_{ki}}$ is given as a function of experimental condition and calculation method. The three calculation methods are structured column wise: the percentile method, the bias corrected method (BC), and the bias corrected and accelerated method (BCa).

Conclusion of analysis 1a and 1b

The finding that the simple percentile method exceeds the more elaborate methods of bias correction, and bias correction and correction for an acceleration parameter astounds on the first glance. However, NMDS analyses, especially analyses that are based on indirect proximities, feature properties which favour confidence intervals calculated by the simple percentile method. We see two sources of variance which hinder adequate results in the more elaborate calculation methods (BC and BCa). First, there is a systematic error which is already introduced at the estimation process of the proximities and second, local minima reduce the reliability of the bias estimators even more.

Firstly, the error from proximity computation is best understood visually, which can be derived from a spatial viewpoint: consider a data structure with a Euclidean metric in, for example, 2 dimensional, standardised space. Suppose that the objects' positions are estimated many times with some degree of error. The distances between the objects vary; occasionally, two objects may even swap their positions in the structure. Given the same error variance it is obvious that two objects located close to each other are more likely to swap their positions than two objects farther apart. However, due to the symmetry property of distances, these changes in space cannot be reflected in the distribution of those distances: the result is a considerably skewed distribution for the estimates of smaller distances and an increasingly better approximation of the normal distribution for the estimates of larger distances. Thus, smaller distances will no longer be normally distributed but will follow a distribution closely linked to the noncentral chi distribution instead. When the error variance about the estimates is exactly 1, the inferred distances actually are chi distributed as was the case in our study in one experimental condition. In this condition, the distances between the cities followed a noncentral chi distribution with noncentrality parameter

$$\lambda_{ij} = \left[\sum_{k}^{n_{y1}} \left(y_{1,i} - y_{1,j}\right)^2 + \sum_{k}^{n_{y2}} \left(y_{2,i} - y_{2,j}\right)^2\right]^{1/2} \quad (6)$$

and degrees of freedom $k = N$. Hence the random error accounts in small distances not only for the variation of the distances but it also systematically increases the distances' expected values. Consequentially, the estimated values must exceed the (transformed) true value, if the number of observations tends to infinity (due to the non-negativity property of distances). Accordingly, the extent of this bias is a function of the distance between two objects, if the error variance is assumed to be constant.

The problem in NMDS analyses is not so much the biased *expected values* of the distances because they still follow monotone transformations (from $\mathrm{E}(\delta_i) > \mathrm{E}(\delta_j)$ follows that $\mathrm{E}(\hat{\delta}_i) > \mathrm{E}(\hat{\delta}_j)$). However, the probability $\mathrm{P}[\hat{d}_i > \hat{d}_j | (d_j > d_i)]$ given $d_j - d_i = \text{const.}$ increases with smaller values of d_i in noncentral chi distributions and introduces a bias into NMDS analyses. As can be seen in Figure 1, chi distributions with smaller noncentrality parameters reveal a larger bias than do distributions with larger noncentrality parameters. The intersections of the vertical lines with the respective distributions in Figure 1 indicate the extent of the bias: if the estimates were unbiased all four intersections would be exactly at a value of 0.5 on the cdf scale. However, it can easily be seen that the intersections occur for smaller distances at smaller cdf values. Moreover, as mentioned above, the bias leads to distorted probabilities $\mathrm{P}[\hat{d}_i > \hat{d}_j | (d_j > d_i)]$ as can be obtained from the smaller area between the curves with noncentrality parameter $\lambda = 1$ and $\lambda = 2$ than from the curves with noncentrality parameters $\lambda = 4$ and $\lambda = 5$.

These biased distance estimates are a methodological flaw that originates already in the extraction of proximity measures from two-way two-mode data. Therefore, the bias is not a bias of the NMDS. It is pre-existing in the raw data i.e. the inferred distance matrix – a fact which has not been considered in NMDS analyses until now to the best of the authors' knowledge.

The second problem for the BC and the BCa method to compute accurate confidence intervals is imposed by local minima in NMDS analyses (Groenen & Heiser, 1996). Local minima could have adverse effects for the computation of the BC and the BCa method: Any difference between $\widehat{\boldsymbol{\theta}}$ and $\overline{\boldsymbol{\theta}}^*$ is inadvertently interpreted as a systematic bias in the BC and the BCa method, even though the difference could have occurred due to a (unsystematic) local minimum. Thus, the introduced correction does not necessarily add to overall precision of the estimate. In our findings, the deviation from the expected α-Level is pronounced in the bias cor-

rected measure already, and thus is likely to be caused by bias estimation. Since there is no systematic difference between the BC and the BCa method, the acceleration constant supposedly plays only a minor role.

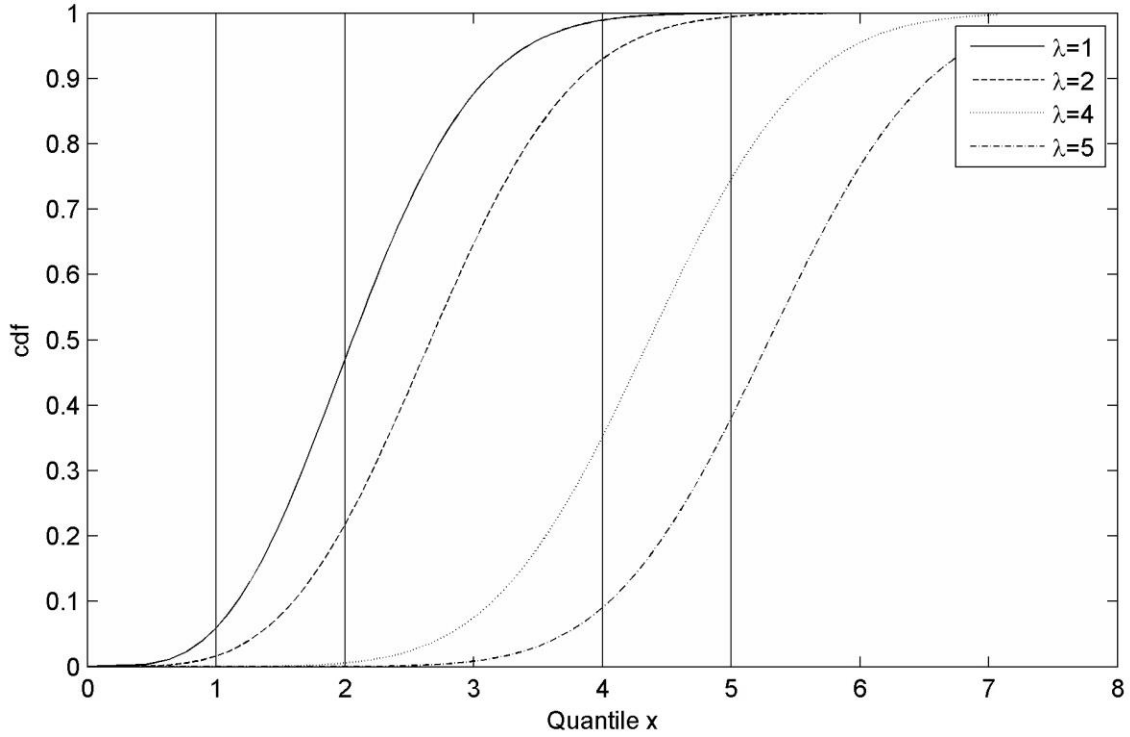

Figure 1. Two groups of chi distributions' ($\lambda = 1, \lambda = 2$ and $\lambda = 4, \lambda = 5$; *df* = 4) cumulative density functions. The vertical lines denote the "true distance" of each distribution. The difference in the areas between the two groups of curves exemplarily reflect the bias in the probability $\mathrm{P}[\hat{d}_i > \hat{d}_j | (d_j > d_i)]$.

Analysis 2

In the previous section we referred to a bias which occurred already when proximities (in our case Euclidian distances) were inferred from two-way two-mode data (indirect proximities). Thus, NMDS analyses are systematically biased, when indirect proximities are used. Additionally, local minima impose a stumbling block for NMDS analysis as Groenen and Heiser (1996) pointed out – especially in low dimensionality. We suggest an approach which addresses both biases: a robust measure of central tendency is inferred from the bootstrap distributions of the NMDS solutions' distances to reduce the influence of local minimum solutions and to correct for biased proximity estimates.

Firstly, the danger of interpreting an NMDS solution that is based on a local minimum solution (and that is structurally different from the "true" solution) can be reduced. There are many suggestions how to deal with local minima in NMDS analyses. Besides improvements in

the NMDS algorithm (Groenen & Heiser, 1996), the fundamental role of the starting configuration has been emphasized (e.g. Borg & Groenen, 2005). Two major approaches to obtain starting configurations are widespread. On the one hand, the starting configuration can be rationally derived, e.g. with classical scaling. Classical scaling assures that the main characteristics of the data are represented quite adequately already at the beginning. On the other hand, a Monte Carlo approach is suggested which re-conducts the NMDS analysis many times with different – e.g. multiple random – starting configurations. Thereafter, the configuration with minimal stress is selected as the global minimum. As promising as the approach of multiple random starting configurations looks, the data specific local minima remain the same in each trial and there is no guarantee that the global minimum or the true configuration was obtained (which must not even necessarily be the same). However, if the data were slightly altered, quite different NMDS solutions might be obtained at the local minima whereas the true structure (regardless if the true structure is found at a global or at a local minimum in the original sample) should be found very similar: after all, the data reveals the true structure in the population. Thus, the true configuration should proof much more robust than any other, randomly occurring local minimum. Of course, the true NMDS solution will not be estimated error free in the altered datasets, but it should occur with only slight structural deviations. With the bootstrapping methods, a tool is already at hand to resample different data sets in accordance with a theoretically sound framework.

Secondly, the robustness property of the central tendency measure benefits estimation of proximities from two-way two-mode data straightforwardly: it prevents the central tendency measure from being heavily influenced by the long tail of the skewed distribution. As a measure of central tendency, we suggest the median because of its innate maximally robust breakdown point of 0.5.

Procedures analysis 2

Of course, a practical implementation could, again, be realised on the level of the distance matrices $\boldsymbol{D}(\widehat{\boldsymbol{\theta}}^*)$ as well as on the level of standardized configurations $\widehat{\boldsymbol{\theta}}_Z$ (as given in eq. 4 & 5). However, we suggest the level of distances, mainly because additional information becomes available for the construction of confidence intervals when distances are used and because analyses 1a and 1b revealed no substantial differences. Hence, the median from the bootstrap distributions was calculated for each distance in the distances matrix ($d_{ij} = \mathrm{Med}\left(d\left(\widehat{\boldsymbol{\theta}}^*\right)_{ij}\right)$). The resulting distance matrix (the medians from the bootstrap distributions) needed to be reanalysed with NMDS, since the matrix did not necessarily satisfy the constraints of a Euclidean space anymore.

In analysis 2, the sum of squared errors of the top level sets were assessed in both, the common NMDS analysis (hereinafter referred to as single step NMDS) and the NMDS analysis,

in which the input data was obtained from the bootstrap distribution (hereinafter referred to as bootstrap based NMDS).

Results analysis 2

Table 5 shows the improvement on the mean squared error of distance $\left(d_{ij} - \hat{d}_{ij}\right)^2$ between single step NMDS and bootstrap based NMDS as a percentage of overall mean squared error and as an absolute value (in brackets). Additionally, the absolute improvement on the mean squared error is shown along with its standard error in Figure 2. The scale of the absolute values in the improvement on the error is arbitrary essentially; but, of course, the same scale is used for the distances in the NMDS solutions. In the software package Prodax (Oberholzer et al., 2008), the distances are normed in a way that the mean of the distances to the centre of mass equals 1.

As can be seen in Table 5, there is an increasing relative effect with increasing sample size. In contrast, the absolute effect decreased with sample size, when the error variance condition remained unchanged. Secondly, the absolute effect increased with increasing error variance when sample size remained unchanged. The relative effect did not show a consistent pattern with varying error variance. In all experimental conditions the bootstrap based NMDS outperformed the single step NMDS analysis with regard to the mean error of distances estimates (Figure 2).

Conclusion of analysis 2

The improvement in the mean of the deviations is quite substantial in relation to the raw deviations and follows tightly the theoretical expectations. The variations of the effect by experimental condition (Table 5) may be explained by the bias from 2W2M data and by local minima. First, the bias from two-way two-mode data should be increasing with increasing error variance (this can be thought of as the ratio between noise (error variance) and information (the distance between two cities which is represented by the noncentrality parameter) in the noncentral chi distribution). Second, the chance of ending up in a local minimum should be depending on both, sample size and error variance, if only because both greatly influence the classical multidimensional scaling solution (which was used to derive the starting configuration). So far, both flaws predict an increase of absolute improvement with increasing error variance. Additionally, the effect of local minima predicts a decrease of absolute effect with increasing sample size due to more precise proximity estimates. Hence, the interaction of these effects presumably accounts for the increase in relative improvement with increasing sample size.

Table 5

Improvement in the mean squared error of distances estimates $\left(d_{ij} - \hat{d}_{ij}\right)^2$ *when the bootstrap based NMDS was conducted instead of the single step NMDS*

	$SD = 0.2$	$SD = 0.5$	$SD = 1$
$N = 50$	10.2% (0.0010)	12.7% (0.0041)	6.4% (0.0083)
$N = 100$	13.7% (0.0008)	12.2% (0.0021)	8.2% (0.0057)
$N = 200$	18.0% (0.0006)	14.7% (0.0013)	13.6% (0.0048)
$N = 400$	21.7% (0.0003)	23.3% (0.0010)	16.0% (0.0029)
$N = 800$	26.5% (0.0003)	29.7% (0.0009)	20.2% (0.0021)

Note. Improvement in the error is given in percentage and as absolute values (in brackets).

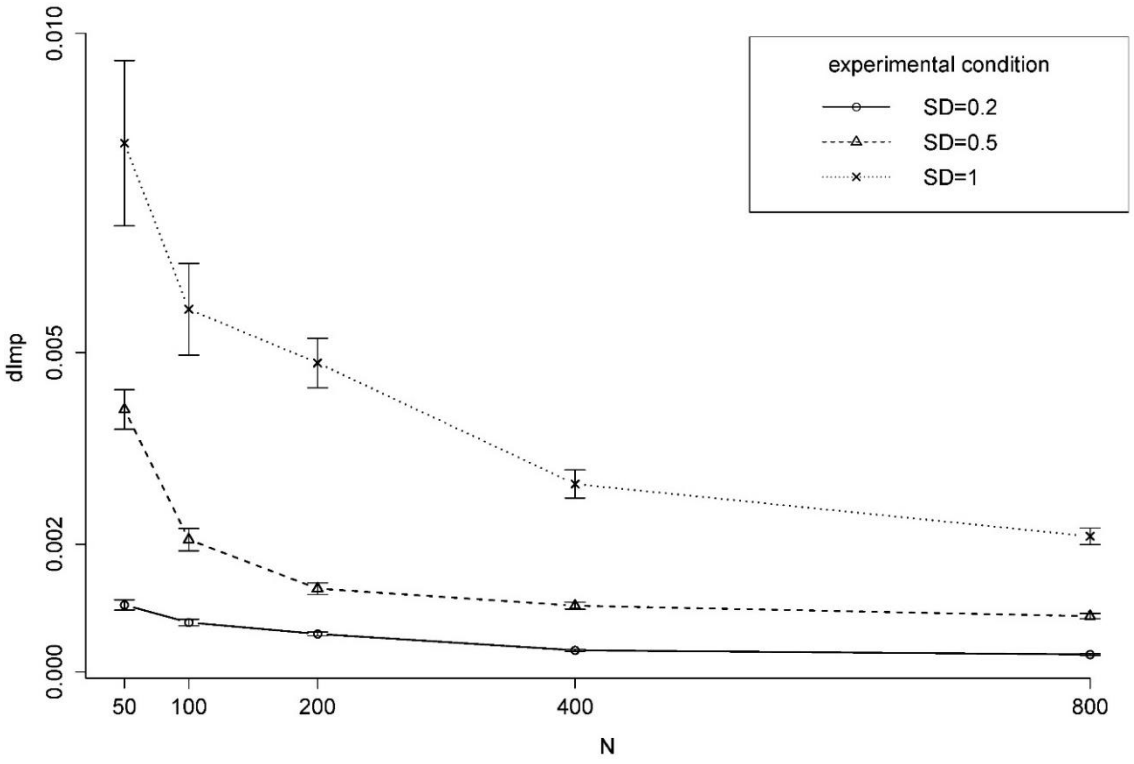

Figure 2. Improvement on the mean squared errors of distances between the cities (*dImp*) when the bootstrap NMDS was used instead of the single step NMDS. The means and standard errors are depicted for the 15 experimental conditions.

Discussion

The present study pursued two main objectives. The first objective was a systematic evaluation of the bootstrap confidence intervals, described by Efron (1979) and Efron & Tibshirani (1986), for NMDS. The second objective was to derive more precise estimates for NMDS analyses by addressing two major deficiencies in NMDS analyses of indirect proximities.

In analysis 1a and 1b confidence regions for the distances and the locations of objects in NMDS solutions were obtained. Three methods (the percentile, the BC and the BCa) were applied in two contexts (distances and locations). A distinct superiority of the percentile method over the more elaborate methods of BC and BCa was found in both contexts.

In analysis 1a, the confidence intervals were calculated based on the bootstrap distribution of distance matrices from bootstrapped NMDS solutions $\boldsymbol{D}(\widehat{\boldsymbol{\theta}}^*)$. The results of the Monte Carlo analysis agreed well with the expected 2α-Level of *0.05* when the percentile method was used to obtain confidence intervals. Surprisingly, when the BC and the BCa methods were applied, the agreement was worse. In analysis 1b the confidence intervals were examined on the level of standardized configurations. Since the dimensional components y_1 and y_2 were treated as independent, we expected a slight overestimation of the correctly included true locations of the objects. The results of analysis 1b were very similar to the results obtained in analysis 1a: the confidence intervals calculated with the percentile method outperformed the BC and the BCa method.

In general, we agree with Weinberg et al. (1984) that the bootstrap produces accurate estimates for confidence regions in NMDS analyses. Additionally, we argued that local minima do not affect the simple percentile method in its accuracy to deliver valid alpha Levels, but that local minima may have a substantial effect on the bias corrected methods. The advanced methods of correction in bootstrapped confidence intervals were likely to produce worse results then the percentile method, if such suboptimal NMDS solution estimates were obtained. The reason for these worse results of the BC and BCa methods is straight forward. A local minimum solution does not deviate from the median of the bootstrap distributions because of a systematic bias of the statistic, but because of an unsystematic error. Hence, the applied correction of the BC and the BCa methods necessarily distort the confidence intervals unsystematically, which presumably accounts for the worse results in Table 2 & Table 4. Furthermore, even though the percentile method may reasonably well produce correct alpha Levels despite local minima, it is evident that the range of confidence intervals may well be reduced by reducing the chance of retrieving a local minimum solution.

Our second aim of this study was the development of an extended estimate procedure for NMDS solutions and was addressed in the second analysis. It was demonstrated in the subsequent analysis that reanalysing the medians from the bootstrap distributions has indeed preferable properties concerning two major deficiencies in NMDS analyses of indirect proximities.

Firstly, the medians from the bootstrap distributions are robust with respect to distortions of proximity estimates and skewed distributions. A systematic bias was pointed out, which occurs when indirect proximities are used: indirect proximities systematically bias the probability $\mathrm{P}\left[\hat{d}_i > \hat{d}_j | \left(d_j > d_i\right)\right]$ given $d_j - d_i = \text{const.}$ if a constant error is assumed on the two-way two-mode data and if symmetric proximity estimates (e.g. distances) are applied. For example in our simulated data, the transformation from two-way two-mode data (with normally distributed errors) to Euclidean distances followed a distribution similar to the noncentral chi distribution. Hence, a substantially skewed distribution and distorted estimators resulted between objects. Secondly, the median is primarily defined by global minima mainly because of its robustness. We argued that (absolute) minima in accordance with the "true" structure are more robust than local minima towards resampling.

The median is a valid estimator for central tendency in such distributions. Its robust properties make it insensitive towards the distributions' long tails and outliers. Even though the proximity bias could be addressed in the computation of distances, we chose to address this issue not before the NMDS was calculated. This procedure allowed us to correct for local minima in the same step. It has to be noted though that the median from the bootstraps does not fully correct for, but certainly reduces the bias in dissimilarity data.

The results from analysis 2 showed a substantial improvement when the median of the bootstrapped NMDS distance matrices was applied instead of the proximity data from two-way two-mode data. The mean squared error in the Monte Carlo analysis towards the "true" structure decreased by 6.4%-29.7%, depending on experimental condition. Additionally, the results were in good theoretical accordance with two biases that depend differently on sample size and error variance. We hypothesized that these error components are: (1) a systematic bias from indirect proximity estimation (independent of sample size but dependent on error variance) and (2) local minima (dependent on both, sample size and variance). Hence, bootstrapping does not only seem to prove useful for calculating confidence regions in NMDS, but also to improve the stability and reliability of NMDS analyses.

The results of both analyses in this study suggest that applying bootstrap methods in an NMDS framework is a promising approach. First of all, it allows testing hypotheses within a structure of objects by calculating confidence regions (e.g. if groups of objects can be considered as homogenous clusters; if objects can be allocated to certain regions within the structure).

Furthermore, it could be shown that estimates from the bootstrap distribution deliver considerably improved configurations. Although not tested, these findings are presumably not restricted to the software ProDax (Oberholzer et al., 2008). The robust weighting function, which is used in ProDax' algorithm RobuScal (Lägeet al., 2005), was disabled specifically for the purpose of maximizing generalizability. Although the beneficial effect of the bootstrap NMDS procedure might decrease to some extent when applied in different Algorithms (e.g. in ProxScal, where the local minimum problem is drastically reduced by a tunnelling algorithm; cf. Groenen & Heiser, 1996), we are confident that it will still outperform the single step NMDS solutions.

Even if the findings apply well to other NMDS algorithms, they should not be thoughtlessly expanded to other domains of NMDS analyses. For example compared to the studies of Heiser and Meulman (1983) or Weinberg et al. (1984), certain differences existed in the input data and in the computation of the bootstrap. These differences complicate the extension of our findings. Among the most prominent methodological differences was the application of indirect proximities instead of direct proximities. It may well be expected, that the error distribution in direct proximities follow other distributions (e.g. the normal or the lognormal distribution), which will presumably diminish the positive effects of the extended NMDS analysis proposed in the current study. Furthermore, a specific setting of NMDS analyses was used, in which the dimensionality of the NMDS solution space was restricted to two and indirect proximities were applied. Additionally, the number of objects was held constant throughout all experimental settings. At last, the "true" configuration was indeed a two dimensional Euclidean structure – which certainly should not be assumed axiomatic for all analysed data (especially in the social sciences). In these respects, our results are limited in generalizability. Nevertheless, evidence was collected that bootstrap methods are well suited to construct confidence intervals in NMDS analyses for a broad bandwidth of sample and variance sizes.

It seems highly plausible that the benefit from using the bootstrap NMDS procedure might increase with a larger number of objects and, vice versa, might decrease with increasing dimensionality. Groenen and Heiser, (1996) reported an increase of local minima when the number of objects was increased and the dimensionality was held constant. Thus, if our hypotheses hold, we expect that the increase of local minima will cause better estimates in the bootstrap NMDS than in the single step NMDS.

Recommendation for application

In consideration of our results from analysis 1a and 1b, we recommend to apply the percentile method instead of the more elaborate BC or BCa methods to compute bootstrapped confidence intervals in NMDS analyses. The results of both analyses revealed good accordance between the theoretical and the empirical alpha level for the percentile method. Regarding the context in which confidence regions were computed, little difference was found between the configuration

(1b) and the distances (1a) context. However, we believe that the distances context is a more adequate level to compute confidence regions and should be applied if feasible.

In a configurations context, the regions are generally calculated as ellipsoids which rely on *k + 1* parameters: one rotational and *k* scaling parameters per ellipse (e.g. Abdi et al., 2009). While this may be a reasonable approach for classical MDS – the classical MDS solution is based on *k* eigenvectors with their respective *k* largest eigenvalues only –, it may well not be successful in NMDS. In NMDS, inter-object distances are not necessarily depending on the same underlying dimensions. The structure is modelled individually on the level of objects. Hence, confidence regions which are based on dimensional parameters only cannot fully account for the objects' different locations. In contrast, in a distances context, an objects α-probability contour (the line which encloses the $1 - \alpha$ confidence region) can be represented as a function in R^k, defined piecewise by the distances $f(d_{i1}, d_{i2}, \dots, d_{ii-1}, d_{ii+1}, \dots, d_{im})$ of their respective bootstrap distributions confidence intervals (where *m* denotes the number of objects). Thus, for our simulated data and NMDS solutions in two dimensions, confidence regions would be the intersections of the according confidence intervals.

An example of such a confidence region is given in Figure 3, where the 95% confidence region for London was constructed for a randomly chosen set of our simulated data ($N = 200$, $SD = 0.5$). As can be seen in Figure 3, not all of the confidence intervals touch the defined confidence region. There are some inconsistencies in the bootstrap distributions that cannot be dissolved in a two dimensional Euclidean space. Nevertheless, we believe this construction method of confidence regions to be more accurate than an elliptical confidence region, though conclusive proof cannot be given in this paper. However, as a matter of practicability, we recommend the computation of ellipsoidal confidence regions in higher dimensional NMDS solutions ($k > 2$). While the construction and interpretation of confidence regions according to our suggestion in two dimensional NMDS solutions is straightforward and inference can be derived visually, higher dimensional solutions complicate these tasks substantially.

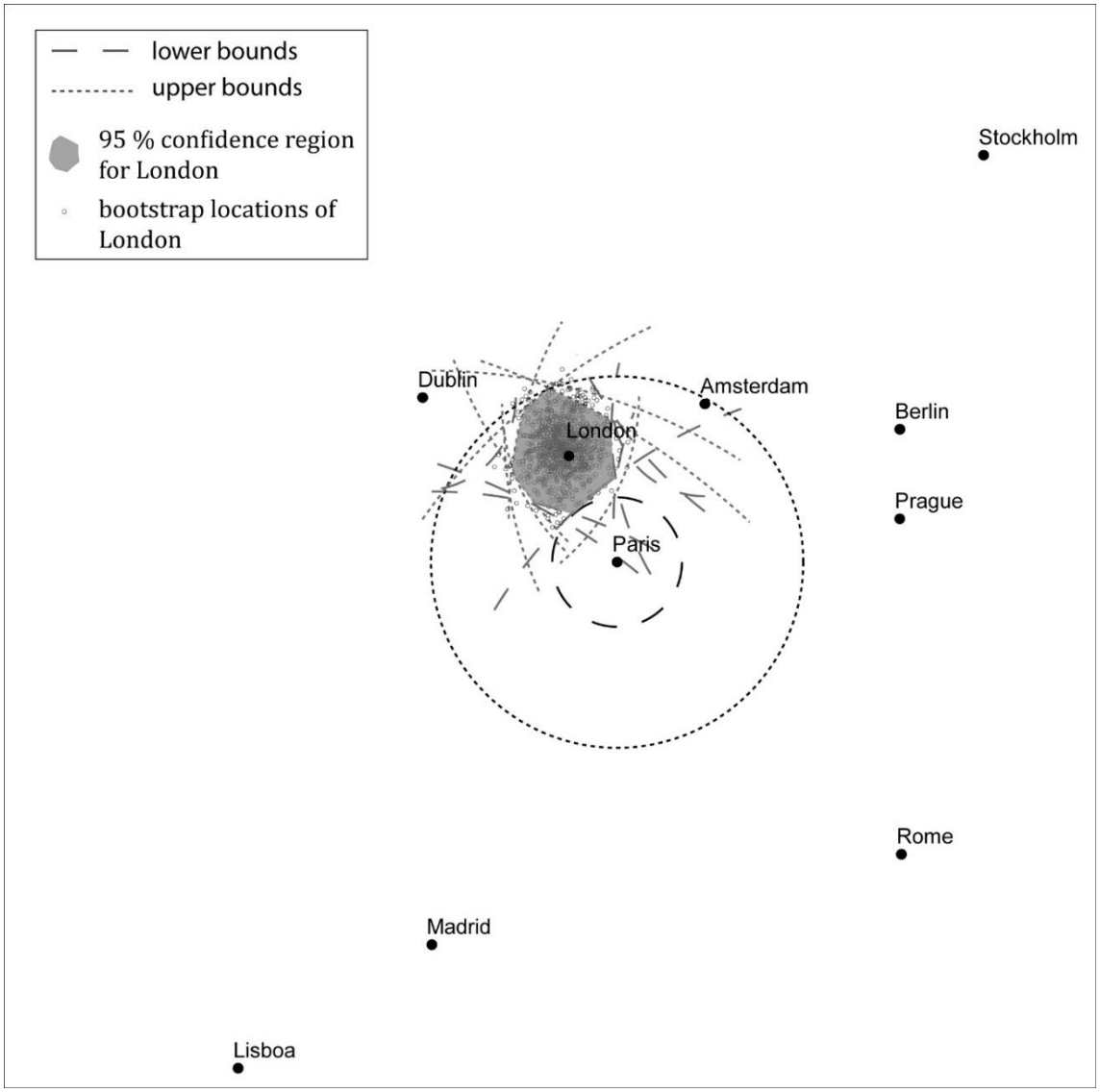

Figure 3. An example for the construction of confidence regions: a random dataset from experimental condition $N = 200$, $SD = 0.5$. For illustrative reasons, the confidence bands are reduced to the region of interest except for London – Paris.

From an applied perspective, these graphical representations of confidence regions impose a major improvement for the interpretation of NMDS results. Firstly, the dissection of an NMDS solution in distinct regions can be achieved (which allows for the grouping of objects). Secondly, the uncertainty of the objects' location estimates in NMDS solutions will become assessable (which allows for hypotheses about underpinned dimensions), and lastly, the expansion of confidence regions allows for an overall estimate on the stability of the result.

These improvements may have a substantial impact on future applications of NMDS. For example Bühler et al. (2012) analysed the symptom structure of the Beck Depression Inventory-II (BDI-II) with NMDS. They identified six facets, which dissected the NMDS solution in six distinct regions. The facets as well as the dimensional ordering of the symptoms were in

good accordance with the theory of depression; however, the stability of the result (i.e. the dissection into regions) could not be properly assessed. Both, the differences between distinct regions as well as the dimensional components of symptom locations could have been assessed if confidence regions were applied.

The number of recent publications with MDS methods and the vast heterogeneity of fields in which these methods are applied document the unbowed significance of MDS among other established methods of analysis (Cox, 2012; Padilla, et al. 2012; Qin, et al. 2012; Vanpoucke, et al. 2012). Correspondingly, the on-going development of the method is essential. We are confident that confidence intervals in NMDS analyses will eliminate some of the reservations towards NMDS results. Moreover we are hopeful that the current findings will inspire others to expand the knowledge and application of the bootstrap in NMDS.

The predictive power of subgroups: an empirical approach to identify depressive symptom patterns that predict response to treatment

Joël Bühler[1], Florian Seemüller[2], Damian Läge[1]

[1]Department of Psychology, University of Zurich, Zurich, Switzerland
[2]Departement of Psychiatry and Psychotherapy, Ludwig-Maximilians-University Munich, Munich, Germany

Submission status:

Submitted to the Journal of Affective Disorders, September 2013.

Authors' contributions:

Joël Bühler: Development of the research question, review of the literature, execution of the analyses, interpretation of the results, writing of the manuscript

Florian Seemüller: Provision of the data, revision of the manuscript

Damian Läge: Supervision and discussion of Joël Bühlers contributions, revision of the manuscript

Abstract

Background: Depression research has been trying to improve the response rates to treatments by identifying a valid set of differential predictor variables. Potential candidates have been proposed, one of which were different subtypes of depression. However, the results on the predictive quality of subtypes on treatment are conflicting.

Methods: The analyzed data consisted of Hamilton Depression Rating Scales (HAM-D_{17}) of 879 depressive inpatients. In a first step, a Latent Class Analysis (LCA) was conducted to classify the patients into smaller groups. In a second step, the class variable was included in a Linear Mixed Effects model to predict the same patients' response to treatment.

Results: Five classes were obtained from LCA, showing substantially different symptom profiles. One of the classes, with a symptom profile similar to melancholic depression, showed substantially slower response to treatment than the remaining classes in the study.

Limitations: The applied measurement instrument, the HAM-D_{17}, did not include items for two additional, frequently found subtypes of depression: psychotic and atypical depression. Thus, these subtypes could not emerge in the LCA. Furthermore, there was no systematic variation of treatment in the data. Thus, a differential effect of the classes on treatment could not be measured.

Conclusions: The classification of patients according to their symptom profiles seems to be a potent predictor for treatment response. However, the obtained symptom patterns are not completely congruent with the theoretically proposed subgroups. Against the background of the results, dividing melancholic depression in a rather cognitive and vegetative subtype may be promising.

Key words: depression, efficacy, subtypes of depression

Introduction

The pursuit of predictors for treatment response in depression research has been going on for decades, yet with only modest success. Both, prescriptive (i.e. predicting differential response to one versus another treatment) and prognostic (i.e. predicting response to a particular treatment) predictors from various domains were identified in the literature (Hollon & Najavits, 1988). However, when looking at the comprehensive meta-reviews (Driessen & Hollon, 2010; Esposito & Goodnick, 2003; Hamilton & Dobson, 2002), most results of the original studies are conflicting. Especially easy to obtain data (e.g. sociodemographic data) have failed to yield good predictions for differential response to treatment (Esposito & Goodnick, 2003).

Nevertheless, some variables have repeatedly been found to influence the response to treatment. With respect to prognostic predictability, Hamilton & Dobson (2002) accentuate the variables high pretreatment severity scores, high chronicity, younger age at onset, an increased number of previous episodes and an unmarried marital status, which seem to be prognostic for poorer response to treatment, at least for CT. Besides pretreatment severity scores, other clinical characteristics have been frequently proposed as predictors of treatment response. Among these clinical predictors, many researchers accentuated the use of different subtypes of depression (e.g. Baumeister & Parker, 2012; Fava, Uebelacker, Alpert, Nierenberg, Pava, & Rosenbaum, 1997). However, the results of studies examining these subtypes as predictors of response to treatment are conflicting as well (Esposito & Goodnick, 2003).

The broad definition of the disorder category “major depression” has raised concerns about its validity as a homogenous category (e.g. Fink & Taylor, 2007; Joiner, Walker, Pettit, Perez, & Cukrowicz, 2005; Lichtenberg & Belmaker, 2010; Stewart, McGrath, Quitkin, & Klein, 2007). Besides concerns about the lack of intrinsic specificity, many of these questions regarding the homogeneity of depression were additionally driven by the high rate of patients with a poor response to treatment and not responding substantially better to medication than to placebo respectively (Turner, Matthews, Linardatos, Tell, & Rosenthal, 2008). Hope remains that treatment response may be drastically increased if an adequate set of depression subtypes could be found, which was able to separate different depressive conditions. Many different subtypes of depression have been proposed, of which a majority were derived from theoretical considerations (e.g. Lichtenberg & Belmaker, 2010; Spitzer, Endicott, & Robins, 1978). However, attempts have also been made to identify subgroups based on various statistical classification techniques (e.g. Aggen, Neale, & Kendler, 2005; Blazer et al., 1989; Cox, Enns, & Larsen, 2001); one of the most promising approaches was Latent Class Analysis (Carragher, Adamson, Bunting, & McCann, 2009; Chen, Eaton, Gallo, & Nestadt, 2000; Eaton, Dryman, Sorenson, & McCutcheon, 1989; Kendler, Eaves, Walters, Neale, Heath, & Kessler, 1996; Sullivan, Prescott, & Kendler, 2002).

To the best of the authors' knowledge, all Latent Class Analyses (LCA) conducted so far were based on samples from the general population and conducted LCA with raw scores or mere symptom occurrences. However, performing LCA with a general population sample and with raw score data yields multiple caveats in the identification of depressive subtypes. First and foremost, LCA should only be conducted with large samples; though if most of the subjects under study do not reveal any depression symptoms at all (e.g. because the sample was drawn from the general population), LCA will not likely obtain symptom patterns, which reliably dissect those few subjects that do reveal depression symptoms. Secondly, depression specific symptom lists (such as the DSM-IV symptom list or any depression specific assessment scale such as the HAM-D) generally yield pronouncedly positively correlated symptoms. The high correlations are usually interpreted as the (unidimensional) severity of depression. However, the classification of patients according to their level of depression severity is not desired if different conditions were to be identified. Instead qualitative differences of the disorder, e.g. the occurrence of different symptom patterns should define the subgroups. Yet, if LCA was performed on raw symptom data, the high correlations between the symptoms would foster the categorization of subtypes according to depression severity levels, overshadowing the more subtle effects of different symptom patterns. We believe the intermingling of depression severity and specific symptom patterns, which resulted in the clinically and theoretically unsatisfactory categorization of depressive patients, is mainly responsible for the little impact of the LCA studies on the concept of depression. This may as well be the cause that none of the studies to date have included statistically derived depression subtypes as predictors for treatment response, even though there is a plethora of literature on the topic.

The aim of the current study was to obtain a typology of depression and to evaluate this typology with respect to its capabilities to predict different response rates to treatment. Therefore, a two-step procedure was applied. In the first step, an LCA was conducted on the patients' mean-centered symptom profiles at baseline (first measurement). The centering ensured the desired independency of the classification from depression severity and the baseline data ensured that the predictor was indeed a pretreatment variable. In the second step, a linear mixed effects model (LMEM) was applied on the course data of the same patients to estimate the effects of subtype on the response to treatment. Both methods, LCA and LMEM, have been applied in depression research before (e.g. Carragher et al., 2009; Fournier et al., 2009). However, a study that combined the two methods had not been conducted, even though it allows direct indications of the usefulness of the obtained typology.

The Hamilton Depression Rating Scale German Version (HAM-D; Collegium Internationale Pyschiatriae Scalarum, 1977) was used to assess the symptom profile of the patients. Even though some authors have raised concerns about its psychometric properties (an overview on the topic is given by Bagby, Ryder, Schuller, & Marshall, 2004), others have pointed out its strengths in additionally assessing symptoms loosely associated with depression (Bech et al.,

1981; Zimmerman, Posternak, & Chelminski, 2005), which, for example, have been found to constitute a strong factor in guiding the selection of antidepressants in treatment (Zimmerman et al., 2004). Furthermore, the HAM-D has endured 50 years of change in the classification of depression (and three major revisions of the DSM!) yet it is still the most widely used clinician administered rating scale for depression. Presumably, it has shaped the way how depression is looked at just as much as did the DSM. Although all patients were assessed with the 21-item version of the HAM-D scale, only the first 17 items were used for the analysis ($HAM\text{-}D_{17}$). We chose to analyze the shortened version of the HAM-D to ensure comparability with previous studies and because the last four items are generally believed to measure depression only poorly (e.g. Hamilton, 1960).

Methods

Sample Characteristics

The sample was collected in a large prospective, naturalistic multicenter study funded by the German Federal Ministry of Education and Research (BMBF). Recruiting of the patients took place at six psychiatric university hospitals and three district hospitals across Germany. Only patients with an age between 18 and 65 were included. Additionally, the diagnostic inclusion criteria required the patients to be diagnosed with a major depressive episode (ICD-10: F31.3x–5x, F32, F33) or with a depressive disorder not otherwise specified (ICD-10: F34, F38, F39) according to ICD-10 (World Health Organization, 1992). The diagnosis was confirmed by the Structured Diagnostic Interview of DSM-IV (SCID; Wittchen, Wunderlich, Gruschwitz, & Zaudig, 1997) and bipolar I and bipolar II disorders were distinguished according to DSM-IV criteria. A total of 1073 patients had been recruited and were, amongst additional rating scales, tested with the 21-item HAM-D scale (Collegium Internationale Pyschiatriae Scalarum, 1977) in biweekly ratings. For the current analysis, additional inclusion criteria had to be met. It was required that each patient had a minimum of two complete HAM-D data sets. The combined inclusion criteria resulted in a reduced set of 879 patients. Of these 879 patients, 62.8% were female and 37.2% were male. The mean age at baseline was 45.1 with a standard deviation of 12.0.

Procedures

The analysis consisted of a two-step procedure. In the first step, an LCA was conducted to divide the sample in homogenous subgroups. The association of the patients to the different classes was then included as a predictor variable for the patients' response to treatment. The second step consisted of a linear mixed effects (LME) model which was applied to the biweekly $HAM\text{-}D_{17}$ ratings. Models of the LME family account for random nested effects due to a hierarchical

data structure and thus were the method of choice to account for the repeated measures with nested random effects in the current data.

LCA procedures to identify symptom patterns

We used the row and column centered HAM-D_{17} symptom scores at baseline to perform the LCA. The row and column centering was applied to eliminate effects of the total score and to ease interpretation of the LCA results, respectively. Subtracting the row mean (row centering) of the data ensured that the patients' total scores of the HAM-D (depression severity) did not influence the LCA results. The column centering was applied for an easier interpretation of the results: hence a mean symptom score of 0 in a given class implied that the class's symptom mean was identical with the grand mean of symptom scores in the total sample. Accordingly, positive values indicated higher mean values of the respective symptoms in the class than in the total sample, whereas negative values indicated lower mean values of the respective symptoms in the class than in the total sample. The column centering had no influence on the calculation of the model. In Contrast, the row centering indeed had implications both on the selection of the LCA model and on the LCA results. The former categorical data was transformed by row centering to (approximately) continuous data; thus a continuous model of LCA was applied.

To compute the LCA, the package MCLUST (Fraley, Raftery, Murphy, & Scrucca, 2012; Fraley & Raftery, 2002) available in the statistical software R (R Development Core Team, 2012) was used, which has been shown to produce reliable results (e.g. Haughton, Legrand, & Woolford, 2009) for continuous data. The 1 to 14 class solutions were calculated and the variance covariance structure was restricted to spherical distributions to restrain the number of parameters needed to estimate the model. The models with less parsimonious variance/covariance structures did not converge, which is a well-known issue and debated in (Fraley et al., 2012), when working with high dimensional data.

Selection of the Latent Class Model. As can be seen in Figure 1, a first peak of the BIC value is reached with five classes, and it decreases again after a plateau of 5 to 8 classes (which all show very similar BIC values), indicating a worse fit for the 9 to 14 class solutions. We considered both, the five (first peak) and the seven classes solution (highest BIC) as possible candidates to categorize the patients into subgroups. As can be obtained from cross tabulation in Table 1, the main characteristics of the seven classes solution is already represented well in the five classes solution, indicating a robust result of classification in the five classes solution. Thus, we chose the five classes solution for further analysis.

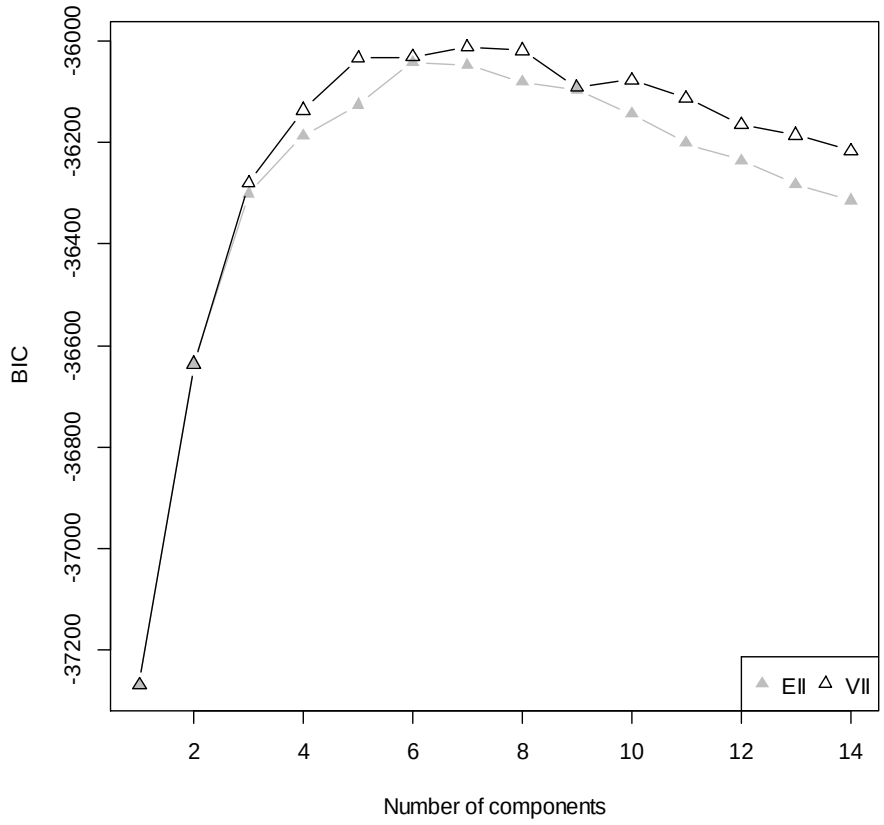

Figure 1. BIC values for different numbers of classes in the centered HAM-D_{17} symptom data. A first peak is reached at five classes. The BIC series VII and EII relate to a variance / covariance structure of variable volume and equal shape, and equal volume and equal shape respectively.

Table 1

Cross tabulation of the five and the seven class solutions

Class	1	2	3	4	5	n (seven class solution)
1	176	8	4	3		191
2	8	15	22			45
3		5	159	1	5	170
4		1		139	8	148
5		112	2	3	8	125
6			1	1	136	138
7	1	30	3		28	62
n (five class solution)	185	171	191	147	185	879

Note. Cells with a grey background indicate the main contingent of patients between the two solutions.

LME procedures

The linear mixed effects model was calculated with the package nlme (Pinheiro, Bates, DebRoy, Sarkar, & R Development Core Team, 2013) available in the software R (R Development Core Team, 2012) and applied to the HAM-D_{17} total score data. For the random effects model, an unstructured covariance structure was applied to model the covariance between the individual intercepts and slopes. Furthermore, maximum likelihood estimation was applied instead of restricted maximum likelihood to be able to assess and compare the fit indices of the models differing in their fixed effects structure (Pinheiro & Bates, 2000).

First, we checked if a model with random linear effects (both intercept and slope) was superior with respect to fit compared to a model with fixed effects only. In the next step, we tried to explain the variability in these random intercept and slope parameters by applying the classification obtained from the LCA. Due to the row centering of the data and the sole use of baseline data in LCA, statistical artifacts due to duplicate application of a subsample of the data (the baseline data) were expected to be minimal.

The classification obtained in the LCA was dummy-coded and included in the LME model as a predictor for the intercept (at baseline) and slope (represented in the model as *class* x *time* interaction). Since the main interest of the current article was the influence of different symptom patterns on response to treatment, the prediction of the slopes was of primary interest.

Results

Results LCA

Table 2 shows the main characteristics of the classification obtained from the LCA of the HAM-D_{17} symptom data. To ease the interpretation of the symptoms' relative importance for the classification solution, a measure related to Cohen's *d* (i.e. $\bar{x}/s$) was chosen to display the symptom score characteristics of the classes. The division of the mean by the symptoms' standard deviation ensured comparability between the symptoms. Furthermore, the preceding column centering of the data ensured that the grand mean of each symptom in the sample was zero.

Two symptoms were especially prominent in the classification of the sample in the five classes: suicide (item 3) and psychic anxiety (item 10). These two symptoms revealed a distinct effect on all five classes, dividing the sample in discriminable subgroups.

Class 1 was characterized by a pronounced, elevated score on suicide and reduced scores on the anxiety related symptoms (items 10 and 11). Patients associated with this class may be best labeled as hopeless.

Table 2

Main characteristics of the five classes obtained in the LCA

HAM-D item	class 1 (hopeless)	class 2 (melancholic)	class 3 (psycho-vegetative)	class 4 (dismayed)	class 5 (anxious)
1. depressed mood	0.32	0.03	-0.26	0.17	-0.15
2. guilt	0.14	0.55*	-0.23	0.20	-0.43*
3. suicide	2.00*	-0.60*	-1.15*	1.07*	-1.26*
4. insomnia, initial	0.29	-0.48*	0.97*	-0.14	-0.51*
5. insomnia, middle	0.17	-0.48*	1.26*	-0.13	-0.53*
6. insomnia, late	0.13	-0.38	0.75*	-0.21	-0.40*
7. work and interests	-0.06	0.49*	-0.43*	-0.06	0.11
8. retardation	0.05	0.81*	-0.45*	-0.30	-0.15
9. agitation	-0.21	0.19	0.04	-0.81*	0.49*
10. anxiety, psychic	-1.14*	-0.94*	-0.54*	1.43*	1.49*
11. anxiety, somatic	-0.79*	-0.27	-0.19	0.60*	0.70*
12. somatic, gastrointestinal	-0.04	-0.13	0.54*	-0.30	-0.12
13. somatic, general	-0.34	0.37	-0.08	-0.12	0.28
14. genital symptoms	-0.09	0.52*	-0.04	-0.18	-0.15
15. hypochondriasis	-0.32	0.49*	-0.03	-0.60*	0.25
16. weight loss	0.07	-0.57*	0.46*	-0.41*	0.21
17. insight	0.12	0.28	0.17	-0.55*	-0.17

Note. To highlight the relative importance of the symptoms for the classification, the values are given as effect sizes with respect to the full data set (i.e. $\bar{x}/\boldsymbol{\sigma}$). * denotes an effect size with an absolute value > 0.4 (medium to large effect sizes according to Cohen, 1988) and was regarded as important for the classification of the respective class.

Class 2 revealed many of the features of melancholic depression, thus it may be best labeled as melancholic. The elevated scores on work and interest (item 7), retardation (item 8), and guilt (item 2) concurred with the features of melancholic depression. However, compared to the other classes, the patients of class 2 did not reveal elevated scores on the psychovegetative symptoms usually associated with melancholic depression: the scores of the insomniac and gastrointestinal symptoms (items 6, 12 & 16) were even slightly below average.

Class 3 was mainly characterized by elevated scores on the three insomnia symptoms (items 4, 5 and 6) and to a lesser degree by the gastrointestinal symptoms (items 12 & 16). The class also revealed reduced scores on the symptoms suicide (item 3) and psychic anxiety (item 10). However, due to the mainly defining cluster of psychovegetative symptoms, we labeled class 3 psychovegetative.

Class 4 revealed substantially elevated scores on the symptoms suicide (item 3), psychic anxiety (item 10) and to a lesser degree on somatic anxiety (item 11). Furthermore, the class revealed reduced scores on agitation (item 9). Class 4 differed from class 1 mainly in the reversed sign on the anxiety symptoms: to reflect this difference in the label, we chose to label the class dismayed.

Lastly, class 5 showed pronounced anxiety symptoms (items 10 and 11) and reduced scores on suicide (item 3) as well as on the insomnia symptoms (items 4, 5, and 6) and guilt (item 2). The reduced scores on suicide (item 3) and guilt (item 2) mainly differentiated the patients of class 5 from the dismayed patients of class 4. Because class 5 revealed substantially elevated anxiety symptoms, the class was labeled anxious.

Results LMEM

The considered models differed in random components and predictor variables. Preceding the main analysis, a log-likelihood ratio test between a general linear model and a random intercepts and slopes model indicated that a model with random effects was indeed superior to a model with fixed effects only ($\chi^2_{diff(3)} = 1596.7, p < 0.001$). Thus, the HAM-$D_{17}$ total score was predicted significantly better by estimating an individual intercept and linear slope parameter for each patient separately.

A second random intercepts and random slopes model was applied, which included the classes as predictors, to test the effect of the predictors (i.e. the dummy-coded class variables) on intercept (HAM-D_{17} score at baseline) and slope (response to treatment). This more specific model revealed improved fit values (AIC/BIC) compared with the model without predictors (Table 3). Thus, the model including the predictors was accepted as a more adequate model. The details of both models can be obtained from Table 3.

The model with predictors revealed significant effects of the classes on both, the HAM-D_{17} score at baseline (main effect of class) as well as on the predicted response (interaction effect of *class* x *time*). Class 2 was chosen as the model baseline, while classes 1 and 3-5 were chosen as contrasts. The negative sign on the *class* x *time* interaction in the model indicated increased response (i.e. faster recovery) of classes 1 and 3-5 compared with class 2. However, the sign of the classes' intercepts were not uniform, indicating lower baseline scores for class 3 and higher baseline scores for classes 4 and 5 compared with class 2. Thus, even though the classes 4 and 5 reveal steeper slopes compared with class 2, it cannot directly be obtained from Table 3, whether this effect also predicts shorter time to remission (HAM-D_{17} total score < 8).

Table 3

Estimated parameters for the random intercepts and slopes models with and without predictor variables from the LCA

	model without predictors			model with predictors		
	γ (SE)		*df*	γ (SE)		*df*
intercept	20.23	(0.20)***	2906	19.73	(0.42)***	2902
time	-1.65	(0.05)***	2906	-1.13	(0.10)***	2902
class 1				0.58	(0.59)	874
class 2				-	(-)	-
class 3				-2.04	(0.59)***	874
class 4				2.38	(0.62)***	874
class 5				2.09	(0.59)***	874
class 1 x *time*				-0.90	(0.16)***	2902
class 2 x *time*				-	(-)	-
class 3 x *time*				-0.39	(0.16)*	2902
class 4 x *time*				-0.64	(0.16)***	2902
class 5 x *time*				-0.72	(0.15)***	2902
AIC	24585.1			24493.7		
BIC	24622.5			24581.1		
-2logLikelihood	24573.1			24465.7		

Note. * $p < 0.05$, ***$p < 0.001$

To determine the differences in time to remission between the classes, the predictive model for each class was solved for the time variable resulting in the following equation:

$$\Delta\hat{t} = \frac{(rem - \beta_0 - \gamma_0)}{(\beta_1 + \gamma_1)}$$

Where $\Delta\hat{t}$ denotes the predicted time to remission, *rem* is the remission score (HAM-D_{17} score of 7), β_0, β_1 denote intercept and slope and γ_0, γ_1 the main effect of *class* and the *class* x *time* interaction respectively. Standard errors for $\Delta\hat{t}$ can be computed by applying Gaussian error propagation. However, calculation of the standard errors are more complex; thus, the detailed formulas are given in the appendix.

Table 4

Estimated time to remission (HAM-D_{17} total score < 8) for the patients in each of the five classes.

Class	n	time to remission (in weeks)	standard error of time to remission
Class 1	185	6.58	1.22
Class 2	171	11.26	1.12
Class 3	191	7.07	1.85
Class 4	147	8.58	1.56
Class 5	185	8.01	1.39

Estimated remission time in weeks is given for each class separately and along with its standard error in Table 4. There is a significant effect of class, indicating longer remission time for class 2 compared with the remaining classes 1 and 3-5 ($Z_i > |2.00|, p_i < 0.05$). None of the pairwise differences in time to remission reached a significant effect for the remaining classes.

Discussion

Discussion of the LCA results

The results obtained from the LCA revealed clinically interpretable and distinct groups of patients, which were mainly characterized by a subset of items and could be labeled as hopeless, melancholic, psychovegetative, dismayed, and anxious. The five class solution was shown to comprise the main characteristics of the seven class solution and thus was preferred due to model parsimony. The good accordance between the five and seven class solutions also suggested that the obtained results were sufficiently stable. The symptom features of the classes reflected those of two frequently found depressive subtypes in the literature (Baumeister & Parker, 2012), although with more detail.

What is described as melancholic/endogenous depression in the literature presumably constituted both, classes 2 (actually labeled melancholic) and 3 (labeled psychovegetative). Thus, patients who revealed mainly vegetative symptoms like insomnia or gastrointestinal symptoms were separated from patients with elevated scores on anhedonia, psychomotor retardation and guilt. The main differentiating characteristic between the two classes may have been a factor associated with the inherent activation level of the symptoms. Class 2 revealed high scores on work and interests (item 7, anhedonia) and psychomotor retardation (item 8), which are both associated with low levels of activation. Contrarily, class 3 revealed substantial below average scores on the two highly activation related symptoms, indicating rather augmented activation. Furthermore, activation looks back on a long standing tradition in the categorization of subgroups in depression research (e.g. Koukopoulos & Koukopoulos, 1999; Shorter, 2007), and in the light of the promising results from the LME analysis, distinguishing the classes two and three might yield positive effects with respect to differential treatment response.

A second, frequently noted subtype of depression in the literature is anxious depression: either described as depression with a comorbid anxiety disorder or, broader, as depression with a generally high level of anxiety (Baumeister & Parker, 2012). Although Baumeister and Parker (2012) refused its qualification as a specific subtype of depression due to possible overlaps with the other three subtypes of depression (i.e. melancholic, psychotic and atypical) and due to the lack of specific treatment effects (e.g. Fava et al., 1997; Lichtenberg & Belmaker, 2010; Rao & Zisook, 2009), the results of this study suggested otherwise. The performed LCA revealed two

classes in concordance with the anxious depression subtype: class 4 (dismayed) and class 5 (actually labeled anxious) were both characterized by substantially elevated scores on the anxiety items (items 10 and 11). The main difference between the two classes was shown on items suicide (item 3) and agitation (item 9) with elevated and reduced scores, respectively, for class 4 and vice versa for class 5. The elevated scores on suicide (item 3) and the reduced scores on agitation (item 9) of class 4 mainly defined the label of the class: "dismayed" – a state in which fear rips away any hope for a turn for the better and paralyzes any counteracting. Thus, the anxiety in class 4 could be described as overwhelming and disabling, while in class 5, it might rather be interpreted as a nervous (highly activated) anxiety.

The classes from the LCA were in good accordance with previous findings of subtypes in depression. In their meta-analysis, Baumeister & Parker (2012) listed four commonly reported subtypes of depression: melancholic, psychotic, atypical, and anxious depression. Two of these subtypes, namely melancholic and anxious depression, were reproduced in the current LCA. However, the subtypes were split in two classes each and thus revealed an increase in detail. The psychotic and atypical depression subtypes were not reflected in the LCA. Please note that these findings should neither be interpreted as a lack of empirical evidence for the existence of these subtypes nor as unreliable results from the performed LCA. Contrarily, the HAM-D_{17} simply does not measure the main characteristics of the psychotic and the atypical subtypes. Hence, those classes could not emerge, given the data from HAM-D_{17}.

Discussion of the LME analysis

In the following LME analysis, a significant main effect of *class* and a *class* x *time* interaction emerged. With the melancholic class (class 2) as baseline class, the dismayed (class 4) and anxious (class 5) classes scored approximately 2 points higher on the HAM-D_{17} scale at admission, whereas the psychovegetative class (class 3) scored 2 points lower. Furthermore, all classes compared (i.e. classes 1, 3-5) revealed significantly steeper slopes, i.e. increased response to treatment (Table 3) and significantly reduced predicted remission times (Table 4). The findings that the melancholic class (class 2) revealed considerably gentler slopes, i.e. showed slower response to treatment on the HAM-D_{17}, is partially in line with previous findings (e.g. Fava et al., 1997), however, others have found no significant effect (e.g. Fournier et al., 2009; Jarrett et al., 2013) as regards the subtypes' prognostic predictions of treatment response. With respect to the melancholic class's prescriptive predictions, it has been noted that Tricyclic antidepressants show better results in treatment than do SSRIs (Esposito & Goodnick, 2003); however, this view has recently been challenged (Driessen & Hollon, 2010).

A possible explanation of the heterogeneous findings on the predicitive abilities of the melancholic subtype in the literature might be explained by the results of the current study. The symptoms commonly associated to depression with melancholic features defined two separate

classes (melancholic and psychovegetative) in the current study. These findings suggest that the melancholic subtype may not be sufficiently homogenous: one group of patients revealed the cognitive symptom complex of melancholia, comprising psychomotor retardation and guilt (class 2); whereas the other group revealed the somatic symptom complex of melancholia comprising weight loss and insomnia (class 3). The results of the LME analysis indicated a distinct effect of slower response to treatment for the melancholic class, whereas the psychovegetative class did not show any differences to the remaining classes (Table 4) with respect to treatment response. However, if the patients of both classes were pooled, a significant effect of response to treatment would unlikely have occurred.

Limitations and general discussion

The main limitations of the study concern the applied measurement instrument. The HAM-D_{17} does not measure the characteristic symptoms of two frequently found subtypes of depression, which are the psychotic and the atypical depression subtypes (Baumeister & Parker, 2012). Furthermore, it was argued that the melancholic class (class 2) and the psychovegetative class (class 3) were both derived from melancholic depression, splitting melancholia into distinct subgroups with cognitive (psychomotor retardation and guilt) and somatic (insomnia and weight loss) symptom complexes. However, based only on the symptoms measured in the HAM-D_{17}, it could not exclusively be determined if these classes satisfy all theoretic needs for being categorized as melancholia; for the symptoms lack of mood reactivity, subjectively different feeling from grief or loss, and worse mood in the morning are not measured in the HAM-D_{17}. Lastly, concerning the study design, all patients included in the study received treatment as usual, due to the naturalistic type of the study. Thus, while yielding excellent external validity, the predictive effect of the classes on response to treatment only allowed for an evaluation with respect to prognostic and not to prescriptive predictions of treatment response.

Nevertheless, the current study is the first to apply LCA on symptom profile data, in which the data was centered previous to the analysis and thus was able to properly disentangle symptom severity and qualitatively different symptom patterns. The lack of previously centering the data led to primarily severity dependent classes in previous studies (Carragher et al., 2009; Chen et al., 2000; Eaton et al., 1989; Kendler et al., 1996; Sullivan et al., 2002), which did not deliver a coherent categorization. The markedly slower response to treatment of the melancholic class (class 2), whose prognostic capabilities were indicated by the LME analysis in the study, might be a potent candidate for differential treatment response. However, future research is needed to replicate the findings of a predominantly cognitive and vegetative subtype of melancholia to confirm the stability of these subtypes. And, most importantly, future research is needed to determine to which extent patients from a cognitive melancholia subgroup might ben-

efit from specific treatment plans. Hopefully, this study will inspire others to keep on researching the different conditions of depression and, eventually, identify those specific treatments clinicians and depressive patients are longing for.

Appendix

Formulas for the estimation of standard errors when solved for time to remission

The general formula for the Gaussian error propagation (with a first order Taylor approximation) is given by

$$u_y = \sqrt{\sum_{i=1}^{m}\left(\frac{\delta y}{\delta x_i}\cdot u_i\right)^2 + 2\sum_{i=1}^{m-1}\sum_{k=i+1}^{m}\left(\frac{\delta y}{\delta x_i}\right)\left(\frac{\delta y}{\delta x_k}\right)\cdot \mathrm{cov}(x_i, x_k)}$$

Where u_y denotes the error of a function $y(x_1, \dots, x_m)$ and u_i denotes the error of x_i.
Applied to the function

$$\Delta\hat{t}(\beta_0, \beta_1, \gamma_0, \gamma_1) = \frac{Rem - \beta_0 - \gamma_0}{\beta_1 + \gamma_1}$$

The standard error of $\Delta\hat{t}$ is estimated for each class (except class 2) as

$$\mathrm{u}_{\Delta t} = \sqrt{\begin{aligned}&\left(-\tfrac{\Delta\beta_0}{\beta_1+\gamma_1}\right)^2 + \left(-\tfrac{\Delta\gamma_0}{\beta_1+\gamma_1}\right)^2 + \left(-\Delta\beta_1\cdot\tfrac{Rem-\beta_0-\gamma_0}{(\beta_1+\gamma_1)^2}\right)^2 + \left(\Delta\gamma_1\cdot\tfrac{Rem-\beta_0-\gamma_0}{(\beta_1+\gamma_1)^2}\right)^2\\ &+2\left[\left(-\tfrac{1}{\beta_1+\gamma_1}\right)\cdot\left(-\tfrac{1}{\beta_1+\gamma_1}\right)\cdot\mathrm{cov}(\beta_0,\gamma_0)\right.\\ &+\left(-\tfrac{1}{\beta_1+\gamma_1}\right)\cdot\left(-\tfrac{Rem-\beta_0-\gamma_0}{(\beta_1+\gamma_1)^2}\right)\cdot\mathrm{cov}(\beta_0,\beta_1)\\ &+\left(-\tfrac{1}{\beta_1+\gamma_1}\right)\cdot\left(-\tfrac{Rem-\beta_0-\gamma_0}{(\beta_1+\gamma_1)^2}\right)\cdot\mathrm{cov}(\beta_0,\gamma_1)\\ &+\left(-\tfrac{1}{\beta_1+\gamma_1}\right)\cdot\left(-\tfrac{Rem-\beta_0-\gamma_0}{(\beta_1+\gamma_1)^2}\right)\cdot\mathrm{cov}(\gamma_0,\beta_1)\\ &+\left(-\tfrac{1}{\beta_1+\gamma_1}\right)\cdot\left(-\tfrac{Rem-\beta_0-\gamma_0}{(\beta_1+\gamma_1)^2}\right)\cdot\mathrm{cov}(\gamma_0,\gamma_1)\\ &\left.+\left(-\tfrac{Rem-\beta_0-\gamma_0}{(\beta_1+\gamma_1)^2}\right)\cdot\left(-\tfrac{Rem-\beta_0-\gamma_0}{(\beta_1+\gamma_1)^2}\right)\cdot\mathrm{cov}(\beta_1,\gamma_1)\right]\end{aligned}}$$

For the baseline class (class 2) the standard error formula reduces to

$$\mathrm{u}'_{\Delta t} = \sqrt{\left(-\frac{\Delta\beta_0}{\beta_1}\right)^2 + \left(-\Delta\beta_1\cdot\frac{Rem-\beta_0}{{\beta_1}^2}\right)^2 + 2\left[\left(-\frac{1}{\beta_1}\right)\cdot\left(-\frac{Rem-\beta_0}{{\beta_1}^2}\right)\cdot\mathrm{cov}(\beta_0,\beta_1)\right]}$$

Where *Rem* denotes remission (a HAM-D_{17} score of 7), β_0, β_1 denote intercept and slope and $\gamma_0, \gamma_1$the main effect of *class* and the interaction of *class* x *time*, respectively.

Berechnung und Interpretation von NMDS Patientenkarten für die Verlaufsdiagnostik: Erste Befunde

[Calculation and interpretation of NMDS patient maps in course diagnostics: first results]

Joël Bühler and Damian Läge

Department of Psychology, University of Zurich, Zurich, Switzerland

Submission status:

Published as a research report in *Forschungsberichte aus der Angewandten Kognitionspsychologie Zürich, 75,* Universität Zürich.

Authors' contributions:

Joël Bühler: Development of the research question, review of the literature, execution of the analyses, development of a correction factor for the NMDS algorithm, interpretation of the results, writing of the manuscript

Damian Läge: Supervision and discussion of Joël Bühlers contributions, revision of the manuscript

Zusammenfassung

In standardisierten psychopathologischen Inventaren werden die Ausprägungen von Symptomen erfasst, die entweder für bestimmte Kategorien von psychischen Störungen oder für die Psychopathologie allgemein bezeichnend sind. Bislang werden diese detaillierten Befunde, die routinemässig bei Patienten erhoben werden, allerdings mit nur wenigen aggregierten Kennzahlen beschrieben, womit ein Grossteil an systematischer Information verlorengeht. Egli, Riedel, Möller, Strauss und Läge (2009) haben das Verfahren der Patientenkarten vorgeschlagen – Eine NMDS-Analyse der psychopathologischen Befunde –, mithilfe derer ein hoher Grad an symptomatischer Detailinformation aus den psychopathologischen Inventaren erhalten bleibt. Die Autoren diskutierten diese Patientenkarten allerdings primär im Hinblick auf die Statusdiagnostik. Im vorliegenden Manuskript wird der Vorschlag von Egli et al. (2009) aufgegriffen und die Möglichkeiten für einen Einsatz der Patientenkarten in der Verlaufsdiagnostik diskutiert. Besonderer Fokus wird auf die methodischen Komplikationen bei der lokalen Interpretation von NMDS-Lösungen gelegt, welche für die Anwendung der Patientenkarten von hoher Wichtigkeit sind. Ein Ansatz zur lokalen Optimierung der Patientenkarten wird aufgezeigt und mit den etablierten NMDS-Algorithmen verglichen. Die lokale Interpretierbarkeit der Patientenkarten bei Verwendung des neuen Ansatzes ist vielversprechend, denn sie übersteigt sowohl in der 2- als auch in der 3-dimensionalen NMDS-Lösung diejenige der etablierten Algorithmen.

Einleitung

Aus der psychiatrischen Status- und Verlaufsdiagnostik sind standardisierte psychopathologische Inventare wie das AMDP (Arbeitsgemeinschaft für Methodik und Dokumentation in der Psychiatrie, 1981), das BDI (Beck, Ward, Mendelsohn, Mock, & Erbaugh, 1961) oder das HAMD (Hamilton, 1960) nicht mehr wegzudenken. Gerade letzteres stellt nach wie vor den „Gold-Standard“ in der Veränderungsmessung bei klinischen Studien mit Antidepressiva dar (Helmreich et al., 2012). Diese standardisierten Inventare bestehen zumeist aus einer Liste psychopathologisch bedeutsamer Symptome, deren Schweregrade separat erfasst wird – entweder durch einen Kliniker oder den Patienten selbst. Die Ausprägungen der einzelnen Symptome werden dann zu Summen aggregiert: Je nach Detailgrad des Instruments zu unterschiedlichen Syndromen (z.B. im AMDP; siehe hierzu Gebhardt, Pietzcker, Strauss, Stöckel, Langer, & Freudenthal, 1983), häufig aber auch zu einem einzigen Gesamtschweregrad einer bestimmten psychischen Störung. Im BDI und HAMD zum Beispiel zum Schweregrad der Depression (Beck et al., 1961). Für die Klassifikation psychischer Störungen sind aber weder die Schweregrade der spezialisierten Inventare (z.B. BDI und HAMD), noch die Syndromwerte der generellen Inventare (z.B. AMDP) ausreichend. In den beiden Diagnosemanualen für psychische Störungen, dem DSM-IV (American Psychiatric Association, 1994) und dem ICD-10 (World Health Organization, 1992), werden die Diagnosekategorien in erster Linie durch Symptomlisten und Auftretensdauer, ganz selten aber durch deren Schweregrad definiert. Aus den aggregierten Massen der psychopathologischen Inventare ist nun allerdings kein Rückschluss auf die darunterliegende Ebene der Symptome mehr möglich: Die Qualität (welche Symptome sind vorhanden) ist untrennbar mit der Quantität (wie stark ist ein bestimmtes Symptom ausgeprägt) der Erkrankung vermischt. Für diagnostische Zwecke, oder gar um Handlungsimplikationen für die Therapie abzuleiten, eignen sich die aggregierten Masse der Inventare entsprechend schlecht.

Allerdings liefern die Diagnosekategorien hinsichtlich der Auswahl therapeutischer Massnahmen ebenfalls keine allumfassende Hilfestellung. So hat Zimmerman (2004) gezeigt, dass die Auswahl der Psychopharmaka für eine Therapie in erster Linie auf dem Vorliegen bestimmter Symptome bzw. Symptomprofile beruht (anstatt auf der Diagnosekategorie), gefolgt von der Vermeidung spezifischer Nebeneffekte und dem Vorliegen von komorbiden Erkrankungen.

An der Wichtigkeit der Symptomatik, sowohl für die Diagnostik als auch für die Behandlung, besteht demzufolge kein Zweifel. Die grosse Schwierigkeit liegt in der Verwertung der symptomatischen Information: Durch die Vielzahl an psychopathologischen Symptomen ist eine schier unerschöpfliche Menge an Symptomkombinationen denkbar, was die Ableitung von Handlungsimplikationen auf der Grundlage von spezifischen Symptomkombinationen verunmöglicht.

Egli et al. (2009) haben ein Verfahren vorgestellt, in dem die Ähnlichkeiten von psychopathologischen Inventardaten (Symptombefunde) mittels Nonmetrischer Multidimensionaler Skalierung (NMDS) abgebildet werden. Sie konnten zeigen, dass sich – allein aufgrund der Symptomprofil-Ähnlichkeiten – die AMDP-Befunde einzelner Patienten der Diagnosen „bipolare affektive Störung, manische Episode ohne psychotische Symptome", „schwere depressive Episode ohne psychotische Episode" und „paranoide Schizophrenie" in 3 distinkte Gruppen ordnen lassen. Die Symptomprofile sind dazu als einzelne Objekte in einen 2-dimensionalen Raum skaliert, in dem die Distanzen zwischen den Objekten deren Ähnlichkeiten repräsentieren: Kleine Distanzen entsprechen einer grossen Ähnlichkeit, grosse Distanzen einer kleinen Ähnlichkeit zwischen den Profilen. Elegant an der Lösung, die Egli et al. (2009) vorschlagen, ist insbesondere die Berechnung der Ähnlichkeiten zwischen den Profilen: Diese wird nicht über die *aggregierten* Masse der Inventare (Gesamt- bzw. Syndromscore) berechnet, sondern direkt über die Ausprägungen der einzelnen Symptome. Sie schlagen dazu zwei unterschiedliche Familien von Ähnlichkeitskoeffizienten vor: Korrelative Ähnlichkeiten und differenzielle Ähnlichkeiten. Damit lieferten Sie eine Lösung für die Integration der symptomatischen Information: Es ist zwar vorderhand unerheblich, durch welche Symptome sich zwei Profile unterscheiden; Profilidentität (als maximale Ähnlichkeit) liegt aber ausschliesslich dann vor, wenn alle Symptome exakt dieselben Werte annehmen. [3] Durch diese Beschränkung der Profilidentität lässt sich aus einer grossen Ähnlichkeitsmatrix (einer Vielzahl an paarweisen Ähnlichkeiten zwischen Objekten) entsprechend auch die gesamte symptomatische Information, praktisch verlustfrei wiederherstellen (nämlich genau dann, wenn die zugrunde liegenden Symptombefunde eine Basis der Symptome bilden). Diese „Rückübertragung" der Ähnlichkeitsmatrix in einen geometrischen (Ziel-)Raum und die Reduktion der Dimensionalität des Zielraums wird durch die NMDS ermöglicht.

Die Befunde von Egli et al. (2009) lassen sich nicht nur in der Statusdiagnostik zukünftig nutzbringend einsetzen; ebenso ist es denkbar, dass die Evaluation von (psychiatrischen bzw. psychotherapeutischen) Behandlungen in solchen *Patientenräumen*[4] geschehen könnte. Denn die fortlaufende Evaluation der Behandlung durch Symptombefunde hat nicht alleinig retrospektiven Charakter. Zwar erlauben die Inventare auch eine objektivere, a posteriori Einschätzung des Behandlungserfolgs bei Psychotherapien (Hannan, Lambert, Harmon, Nielsen, Smart, & Shimokawa, 2005) – was für die Qualitätssicherung grosse Bedeutung hat. Weitaus wichtiger aber sind Befunde zur Wirkung von Feedback auf den Behandlungsprozess selbst (Slade, Lambert, Harmon, Smart, & Bailey, 2008): Mittels Feedback zum Zustand des Patienten konnten

[3] Die Beschränkung auf eine einzige Identität gilt bei differenziellen Massen. Bei Korrelationen sind die Einschränkungen für Identität etwas geringer. Bei der Pearson Korrelation zum Beispiel die Menge aller Z-Transformationen der gegebenen Befundausprägungen.

[4] Die Patientenräume bezeichnen dabei die NMDS-Lösungen auf der Basis der psychopathologischen Befunde.

wesentlich bessere psychotherapeutische Behandlungsergebnisse erzielt werden, als in Behandlungen, bei denen kein Feedback gegeben wurde. Der Effekt in der Studie von Slade et al. (2008) zeigte sich bei einem wöchentlichen Feedback zum Zustand des Patienten. Das Feedback umfasste dabei sowohl Informationen zum Schweregrad als auch auffällige bzw. klinisch besonders relevante Symptome wie Suizidalität.

Der Nutzen von Patientenkarten, welche ein ausserordentlich detailliertes Feedback zum Zustand des Patienten geben könnten, dürfte sich demnach nicht allein auf die Statusdiagnostik beschränken, sondern könnte insbesondere auch in der Verlaufsdiagnostik und Behandlungsplanung vorhanden sein. Um ein valides Instrument für die Verlaufsdiagnostik einer Behandlung zu entwickeln bedarf es allerdings vorgängig einer differenzierten Auseinandersetzung mit der Datenlage und der Auswertemethodik, um mögliche methodischen Artefakte auszuschliessen bzw. darauf hinzuweisen. Das Ziel des vorliegenden Artikels ist der erste Schritt in diese Richtung: Er soll eine Übersicht liefern über die bekannten und noch nicht publizierten Hindernisse auf dem Weg hin zu einer methodisch sauberen und praxistauglichen Patientenkarte.

Notwendigerweise mussten einige Einschränkungen der Allgemeinheit vorgenommen werden, um den Umfang des Artikels auf überschaubarem Niveau zu halten. Die Wichtigste ist wohl die Beschränkung auf unidimensionale Inventare, denn die hierin präsentierten empirischen Befunde stammen allesamt aus der Analyse von BDI-Befunden. Da es sich beim vorliegenden Manuskript nicht um eine klassisch empirische Studie sondern im Hauptteil um theoriegeleitete Überlegungen handelt, wurde folgende Abschnittsstruktur gewählt:

- Berechnung von Patientenkarten für die Verlaufsdiagnostik
- Der Behandlungsverlauf in den Patientenkarten
- Inverse-Interpretation: Ähnlichkeiten von Nachbarobjekten in NMDS-Lösungen
- Der Einfluss der Stresswert-Funktion und der Dimensionalität auf das Inverse-Interpretationsproblem in Patientenkarten
- Diskussion

Während der erste Abschnitt „Berechnung von Patientenkarten für die Verlaufsdiagnostik“ in erster Linie die methodischen Eigentümlichkeiten und Voraussetzungen sowie interpretative Leitfäden für die Patientenkarten fokussieren, ist der zweite Abschnitt, „Der Behandlungsverlauf in den Patientenkarten“, spezifisch auf die klinische Anwendbarkeit der Patientenkarten ausgerichtet. Im dritten Abschnitt wird das Problem der „Inversen-Interpretation“ – dem Rückschluss von Distanzen aus der NMDS-Lösung auf die Ähnlichkeiten der Befunde – thematisiert und ein algorithmischer Ansatz zur Optimierung des Problems skizziert, zu welchem im vierten Abschnitt die ersten empirischen Befunde präsentiert werden.

Trotz der primär theoretischen Ausrichtung dieses Artikels wird in den folgenden Abschnitten immer wieder auf Abbildungen rekurriert, deren Inhalte mehr als nur symbolischen Charakter haben. Zwar werden, bis auf den letzten Abschnitt, keine streng quantitativen Analysen präsentiert; um den theoretischen Nachvollzug zu erleichtern, wird aber bereits vor dem letzten Abschnitt ab und an auf klinisches Datenmaterial zurückgegriffen. Die Analysen und Illustrationen basieren durchwegs auf derselben Datenbasis, weshalb diese bereits an dieser Stelle erläutert wird.

Sample Charakteristika

Der Datensatz bestand aus BDI-Daten unipolar depressiver Patienten (N = 85), welche an der Clienia Schlössli (stationär) im Rahmen der Zürcher Stufenplanstudie (Montani, 2009) erhoben wurden. Neben dem BDI wurden zusätzlich weitere psychopathologische Inventare und biologische Marker erhoben, die im weiteren Verlauf der Analyse keine Relevanz besitzen. Die Daten wurden zwischen Ein- und Austritt wöchentlich erhoben, wodurch sich die Anzahl Zeitpunkte je Patient nach der Dauer seiner individuellen Behandlung richtete. Von den 85 Patienten wurden nur diejenigen in die vorliegende Analyse einbezogen, die eine protokollkonforme Behandlung durchlaufen hatten und mindestens einen BDI Befund aufwiesen. Dies führte zur Reduktion auf 45 Patienten. Das Durchschnittsalter der eingeschlossenen Patienten lag bei 42.4 Jahren (SD = 11.3), der Frauenanteil betrug 57.8%. Die durchschnittliche Behandlungsdauer betrug 8.0 Wochen (SD = 3.35). Für eine vollständige Beschreibung des Datensatzes inklusive ausgeschlossener Inventare und biologischer Marker wird auf die Arbeit von Montani (2009) verwiesen.

30 Patienten wurden zufällig als Referenzsample ausgewählt. Bei den Patienten des Referenzsamples wurde jeweils deren Ein- und Austrittsbefund (insgesamt 60 Befunde) einbezogen. Bei den verbleibenden 15 Patienten wurden alle Befundzeitpunkte einbezogen (insgesamt 72 Befundzeitpunkte).

Berechnung von Patientenkarten für die Verlaufsdiagnostik

Entgegen dem in Egli et al. (2009) verwendeten korrelativen Ähnlichkeitsmass, scheint für die Verlaufsdiagnostik ein differenzielles Ähnlichkeitsmass (bzw. Distanzmass) zweckmässiger: Während bei korrelativen Ähnlichkeiten der Gesamtscore (Schweregrad) der Profile keinen Einfluss auf deren Ähnlichkeiten ausübt, beeinflussen die Gesamtscores die Ähnlichkeiten bei Distanzmassen erheblich. Da bei Behandlungen von psychischen Störungen die Reduktion des Schweregrads eine sehr zentrale Rolle einnimmt, sollten Mittelwertsunterschiede zwischen verglichenen Symptomprofilen (Schweregradunterschiede) auch entsprechend in die Ähnlichkeitskoeffizienten einfliessen. Eine Familie von differenziellen Ähnlichkeitskoeffizienten sind die Minkowski-Distanzen. Sie sind gegeben durch

$$d_{ij} = \left(\sum_{a}^{m} \left| x_{ia} - x_{ja} \right|^{p} \right)^{\frac{1}{p}} \quad (1)$$

wobei m der Dimensionalität – also der Anzahl der Symptome eines Inventars – entspricht. x bezeichnet ein beliebiges Symptomprofil und p bestimmt die spezifische Art der Metrik, wobei grundsätzlich beliebige Wert denkbar sind. Die am häufigsten verwendeten Metriken sind $p = 2$ (euklidische Metrik) und $p = 1$ (City-Block Metrik). Generell gilt: Je grösser p, desto grösser der Einfluss von „grossen" Differenzen. Das heisst, die Distanz (d_{ij}) wird mit ansteigendem p immer stärker durch die grösste Differenz bestimmt, bis zum Extremfall der Dominanzmetrik (bezeichnet mit $p = \infty$), in dem $d_{ij} = \max(\left| x_{ia} - x_{ja} \right|)$. In den Patientenkarten im vorliegenden Manuskript wird durchgängig die City-Block Metrik verwendet, da a priori keine theoretische Rechtfertigung vorliegt, grossen Differenzen einzelner Symptome überproportionales Gewicht zu verleihen.

Die Ähnlichkeiten zwischen je zwei Symptombefunden werden entsprechend als Summe der Beträge ihrer symptomatischen Unterschiede erfasst. Aus einer Gruppe von N Symptombefunden entstehen so $\frac{N \cdot (N-1)}{2}$ paarweise Ähnlichkeitsbeziehungen, die schlussendlich mittels NMDS in einen niedrig dimensionalen Raum abgebildet werden (Abbildung 1). Da ein differenzielles Ähnlichkeitsmass gewählt wurde, verhalten sich die Ähnlichkeiten zu den Distanzen in der NMDS-Lösung umgekehrt proportional. Generell gilt entsprechend: Je grösser die City-Block Distanz zwischen zwei Befunden, desto grösser die Distanz zwischen den Befunden in der NMDS-Lösung.

In NMDS-Lösungen ist die Dimensionalität gegenüber den Ursprungsdaten i.d.R. stark reduziert. In Abbildung 1, in der die NMDS-Lösung auf der Grundlage von BDI-Daten berechnet wurde, wurde die Dimensionalität von ursprünglich 21 (der BDI besteht aus 21 Symptomen) auf 2 reduziert. Der Vorteil der Dimensionsreduktion liegt in der besseren Interpretierbarkeit der Struktur (denn so werden die wichtigsten Unterschiede betont), der Nachteil liegt offenkundig im Verlust von Detailinformation; denn durch die Dimensionsreduktion lässt sich der Informationsgehalt der Daten nicht mehr verlustfrei abbilden. Für die Transformation der (Un-)Ähnlichkeiten (den City-Block Distanzen aus den Symptomprofilen) in euklidische Distanzen der Zieldimensionalität sucht die NMDS den am besten passenden Kompromiss. Die Transformation erfolgt dabei möglichst rangtreu: Eine Lösung gilt also auch dann noch als perfekt, wenn zwar die ursprünglichen Intervalle verletzt, aber keine Rangplatzverschiebungen zwischen den Disparitäten (den ordinal transformierten Unähnlichkeiten) und den Distanzen aus der NMDS vorliegen. Der beste Kompromiss wird durch die iterative Minimierung der Stresswert-Funktion erreicht (Borg & Groenen, 2005), wobei sich die Funktion je nach Algorithmus unterscheidet. Im vorliegenden Manuskript wird in erster Linie der Algorithmus RobuScal verwendet, der im

Softwarepaket ProDax (2008) implementiert ist und dessen Stresswert-Funktion gegeben ist durch:

$$\sigma^2(X) \coloneqq \frac{\sum_{j<i} w_{ij}\left(\delta_{\mathrm{ij}} - d_{ij}\right)^2}{\sum_{i<j} w_{ij} d_{ij}^2} \qquad (2)$$

Darin bezeichnet δ die Disparitäten und d die Distanzen in der NMDS-Lösung. RobuScal unterscheidet sich in erster Linie durch den Einbezug einer Gewichtungskonstante (w) in der Stresswertfunktion von den übrigen Stresswertfunktionen, was diese robuster gegenüber grossen Fehlern macht (Läge, Daub, Bosia, Jäger, & Ryf, 2005).

NMDS Lösungen sind immer relationale Abbildungen einer Datenstruktur. Die Distanzen in einer NMDS Lösung sind entsprechend nur innerhalb dieser spezifischen Lösung interpretierbar. Ändert sich der Datensatz, ändert sich also immer auch die Interpretation der „Länge“ einer Distanz – auch wenn es sich um eine Distanz zwischen denselben Objekten (in unterschiedlichen Datensätzen) handelt. Der Effekt bei nur wenigen veränderten Objekten und ähnlicher Varianz in der Ähnlichkeitsmatrix ist freilich gering. Trotzdem sollte der Effekt beim Vergleich von NMDS Lösungen immer mit berücksichtigt werden.

Patientenräume, denen Daten von unidimensionalen Inventaren (z.B. BDI-Daten; Abbildung 1) und differenzielle Ähnlichkeiten zugrunde liegen, zeigen meist ein spezifisches Phänomen: Bildlich gesprochen nimmt die Punktwolke der Befunde in der NMDS Lösung die Form eines Kometen an. Auf der einen Seite (in Abbildung 1 rechts) finden sich, dicht gedrängt, eine Vielzahl an Austrittsbefunden, auf der anderen Seite (in Abbildung 1 links) dagegen, breit gestreut, vorwiegend Eintrittsbefunde. Die Punktwolke in Abbildung 1 kann so als Meteor (rechts) mit breiter werdendem Schweif (links) gesehen werden.

Die Ausrichtung dieser kometenhaften Form folgt zumeist der x-Achse und repräsentiert die Unähnlichkeiten der Symptomprofile aufgrund divergierender Gesamtscores in den Profilen. Da bei unidimensionalen Inventaren i.d.R. durchwegs hohe interne Konsistenzen zu finden sind und sowohl Eintritts- als auch Austrittsbefunde im Referenzsample verwendet werden, zeigt sich die Hauptvariabilität zwischen den Profilen primär als Schweregradabhängigkeit. Denn bei hohen internen Konsistenzen zeigen die Symptome generell auch hohe Korrelationen untereinander – die Wahrscheinlichkeit ist also höher, dass ein Symptom hohe Ausprägungen zeigt, wenn die übrigen Symptome ebenfalls hohe Ausprägungen annehmen, als dass es tiefe Ausprägungen zeigt, bei hohen Ausprägungen der übrigen Symptome. Obwohl durch die City-Block Distanz theoretisch dieselbe Variabilität durch Profilunterschiedlichkeit (gegenläufige Symptomausprägungen, gleicher Schweregrad) wie durch Schweregradunterschiedlichkeit (gleiche Profile, unterschiedlicher Schweregrad) resultieren könnte, ist dies in der Praxis nur bei extra zu diesem Zweck konstruierten Referenzsamples oder eventuell bei multidimensionalen Inventaren der Fall.

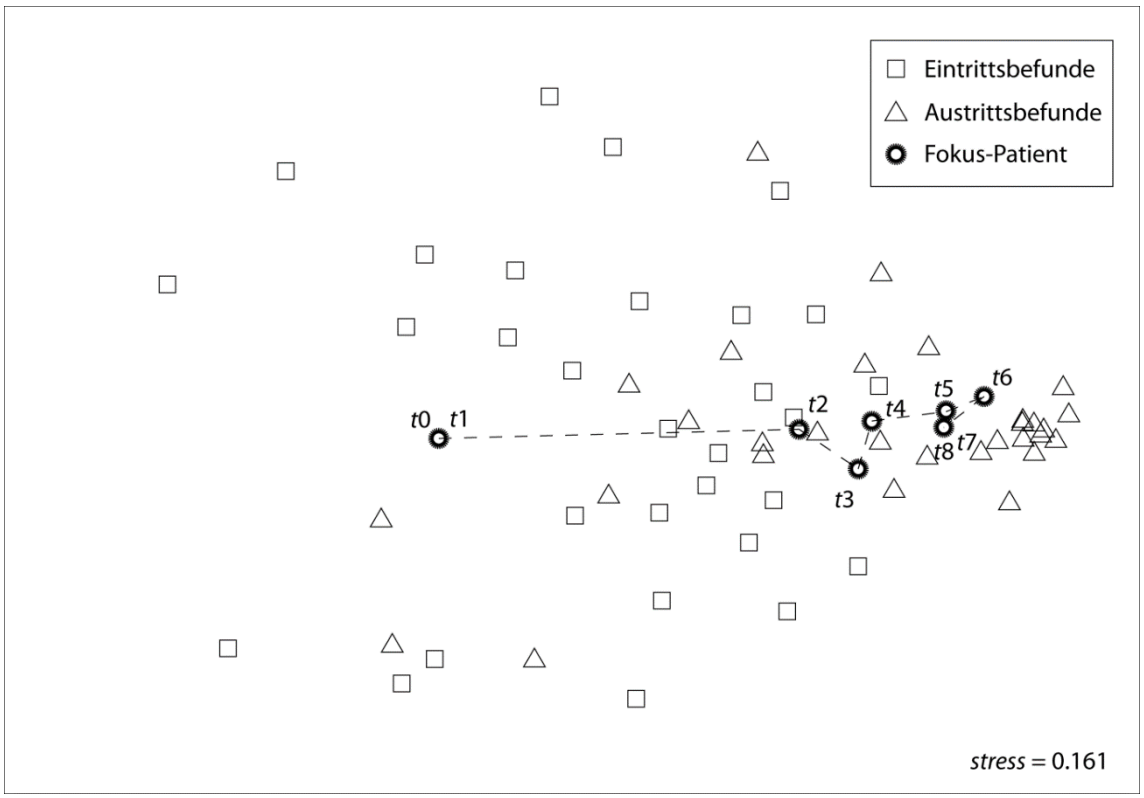

Abbildung 1. Patientenkarte mit 30 (Referenz-) + 1 (Fokus-) Patienten. Jeder Referenzpatient ist mit zwei (Eintritt & Austritt), der Fokuspatient mit 9 Symptomprofilen (Eintritt, Zwischenbefunde & Austritt) vertreten. In zwei Zeitintervallen (t_1-t_2 & t_8-t_9) hat keinerlei Veränderung in der Symptomatik des Fokuspatienten stattgefunden.

Die unterschiedliche Streubreite in der Vertikalen zwischen Profilen mit hohem und mit niedrigem Schweregrad lässt sich auf die systematische Heteroskedastizität zwischen Profilen gleichen Schweregrads zurückführen. Befunde mit geringem Schweregrad können sich nur unwesentlich voneinander unterscheiden, da sie per Definition nur wenige Ausprägungen aufweisen, die Varianz in den Ähnlichkeiten ist entsprechend begrenzt. Dies führt automatisch zu starker Clusterung dieser Befunde in den Patientenkarten, was sich auf der rechten Seite von Abbildung 1 zeigt (der Meteor). Im Gegensatz dazu lassen höhere Schweregrade, insbesondere Schweregrade im mittleren Bereich, eine grössere Schwankungsbreite der Profilunähnlichkeit bei gleichem Schweregrad zu. Diese zusätzliche Variabilität kann nicht entlang der Dimension der Schweregrade (in Abbildung 1 entlang der x-Achse) abgebildet werden. Entlang der zweiten Dimension, in Abbildung 1 also entlang der y-Achse, werden entsprechend massgeblich die qualitativen Unterschiede in den Symptomprofilen modelliert (der breiter werdende Schweif). Theoretisch müssten Profile mit den allerhöchsten Schweregraden ebenfalls wieder sehr stark clustern, da sie, genau wie Profile mit sehr geringem Schweregrad, nur unwesentlich Variabilität im Symptomprofil aufweisen können. Praktisch werden diese allerhöchsten Schweregrade aber ganz selten erreicht, weshalb eine Clusterung auf der gegenüberliegenden Seite zum Austrittscluster nur in einem hochgradig konstruierten Setting zu erwarten ist.

Die spezifische Ausrichtung der meisten Patientenkarten entlang der x- (Schweregrad) und y-Achse (qualitative Unterschiede) liegt nicht etwa in der Lösung selbst begründet, denn NMDS Lösungen sind invariant gegenüber Rotation, Translation, Spiegelung und Skalierung. Vielmehr widerspiegelt die Ausrichtung der Lösung die verwendete Startkonfiguration: In RobuScal wird für diesen Zweck die klassische Skalierung (eine Art Hauptkomponentenanalyse) verwendet. Bei einer zweidimensionalen NMDS werden die Faktorladungen der ersten zwei Hauptkomponenten als Startkonfiguration gesetzt, was die Hauptvariabilität im Datensatz bereits recht gut repräsentiert. Während der NMDS ändern sich zwar die Positionen der Objekte, die Ausrichtung der Hauptkomponenten bleibt aber meist erhalten. Dies führt dazu, dass die Hauptkomponenten, auch in der finalen NMDS Lösung, weiterhin nach der x- und y-Achse ausgerichtet bleiben.[5]

Der Behandlungsverlauf in den Patientenkarten

Abbildung 1 zeigt eine Patientenkarte, in welcher der Behandlungsverlauf eines interessierenden Patienten (Fokuspatienten) bereits grafisch markiert wurde (die Behandlung beginnt links und folgt der gestrichelten Linie). Die Interpretation des Behandlungsverlaufs in den Patientenkarten ist hochgradig intuitiv, denn sie wird direkt als Bewegung des Patienten im Patientenraum repräsentiert. Aufgespannt wird der Patientenraum durch ein Sample von Referenzpatienten, deren Symptomprofile zwar nicht direkt interessieren, die aber für die Interpretation des Behandlungsverlaufs von äusserster Wichtigkeit sind. Im Falle von Abbildung 1 sind zusätzlich zum Fokuspatienten 30 Referenzpatienten jeweils mit deren Ein- und Austrittsbefund in der Patientenkarte vertreten. Insgesamt sind also 60 Symptomprofile vorhanden, die massgeblich den Patienraum definieren. Für die Verlaufsdiagnostik bei psychischen Erkrankungen werden durch die Relationalität der NMDS Lösungen spezifische Anforderungen an die Referenzdaten gestellt: Idealerweise wird sowohl die Variabilität des Schweregrads als auch die qualitative Variabilität in der Symptomatik, welche in den Symptomprofilen des Fokuspatienten vorhanden sind, durch das verwendete Referenzsample abgedeckt. Damit wird gewährleistet, dass sich innerhalb der Referenzgruppe überhaupt vergleichbare Befunde zum Fokuspatienten finden lassen.

Der Fokuspatient in Abbildung 1 zeigt insgesamt stetige Besserung ohne substantielle Veränderungen in der symptomatischen Qualität (Symptomverschiebungen) seiner Erkrankung. Die vorwiegend horizontal verlaufenden Verbindungslinien zwischen den Zeitpunkten $t_1 - t_8$ zeigen zumeist eine Reduktion im Schweregrad an: Besonders deutlich war die Veränderung

[5] Im Folgenden wird davon ausgegangen, dass die Patientenkarten immer derart gedreht werden, dass in der Horizontalen die Variabilität der Schweregrade und in der Vertikalen die Variabilität der symptomatischen Qualität aufgespannt wird. Durch die Invarianz von NMDS Lösungen gegenüber Rotation ist eine derart standardisierte Lösung mit der Ursprünglichen identisch und dient einzig dem einfacheren Referenzieren in den verbleibenden Passagen.

zwischen Zeitpunkt t_1 und Zeitpunkt t_2, was spontan an die von Tang und DeRubeis (1999) dokumentierten „Sudden Gains“ erinnert; diejenigen ausserordentlichen Behandlungsfortschritte zwischen zwei Therapiesitzungen, die für rund 50 Prozent der Gesamtbesserung während der Behandlung verantwortlich sind. Dass es sich bei der Veränderung zwischen t_1 und t_2 um eine Schweregrad*reduktion* handelt, lässt sich an der Position zum Zeitpunkt t_2 ersehen. Diese ist wesentlich dichter am Cluster der Austrittsbefunde dran als noch zum Zeitpunkt t_0 und t_1. Die Bewegung erfolgt also auf das Cluster der Austrittsbefunde zu, weshalb es sich um eine Reduktion des Schweregrads handeln muss. Vertikale Verschiebungen der Fokuspatientenbefunde, was Symptomverschiebungen entsprechen würde, liegen zu keinen Zeitpunkten deutlich vor. Einige geringe Verschiebungen zwischen den Zeitpunkten t_2 und t_4 sind, im Vergleich zu den drastischen Schweregradreduktionen, kaum erwähnenswert.

Vergleich von Fokuspatienten-Verläufen.

Neben der individuellen Analyse eines Fokuspatienten ist es durchaus denkbar, dass die Behandlungsverläufe mehrerer Fokuspatienten miteinander verglichen werden sollen. Dazu könnten die Profile unterschiedlicher Fokuspatienten, zusammen mit dem immer gleichen Referenzsample, separat mit NMDS skaliert werden. Aus den (standardisiert gedrehten) Patientenräumen liessen sich so standardisierte x- und y- Koordinaten der Fokuspatientenbefunde zu den verschiedenen Zeitpunkten extrahieren. Neben der Schweregradveränderung (die sich natürlich wesentlich einfacher direkt aus der Symptomsumme der Fokuspatientenbefunde ableiten liesse) könnte zusätzlich ein Mass der symptomatischen Veränderungen während der Behandlung gewonnen werden (die y-Koordinaten) und zwar in direkter Relation zum verwendeten Referenzsample.

Durch den Austausch der Fokuspatientenbefunde ändert sich allerdings zwangsläufig der skalierte Datensatz, womit auch Veränderungen des Patientenraumes einhergehen werden. Dies bringt eine etwas unschöne Beeinflussung des Messsystems durch den Messgegenstand mit sich, dessen Relevanz auf dem jetzigen Stand noch nicht vollumfänglich abgeschätzt werden kann. Erste Erfahrungen zur Stabilität der Schweregraddimension liegen zwar vor: Die Korrelation der x-Werte der Fokuspatientenbefunde mit den zugehörigen Schweregraden im vorliegenden Datensatz erreichte mit $N = 15$ Patienten bei 72 Zeitpunkten nahezu 1 ($r = 0.988$). Dieses Resultat legt bezüglich des Schweregrads in den NMDS Lösungen Skaleninvarianz nahe, trotz immer leicht unterschiedlicher Datenbasis (gleiche Referenzbefunde, andere Fokuspatientenbefunde). Denn angenommen die Metrik in den NMDS Lösungen würde sich durch die veränderte Datenbasis der Fokuspatienten massgeblich verändern, dann müssten dieselben Schweregrade zu substantiell unterschiedlichen Skalenwerten in der Patientenkarte führen, was durch die Höhe der Korrelation als äusserst unwahrscheinlich angesehen werden darf.

Es gilt jedoch zu bedenken, dass der Schweregrad in Datensätzen unidimensionaler Inventare allerdings die mit Abstand wichtigste Hauptkomponente darstellt, was eine Verallgemeinerung auf die qualitative Dimension (die y-Achse) erschwert. Denn erstens reicht die Stabilität der qualitativen Dimension kaum an diejenige der Schweregraddimension heran (bei der Schweregraddimension wird durch den Einbezug von Ein- (hoher Schweregrad) und Austrittsbefunden (niedriger Schweregrad) eine stabile Verankerung gegeben). Und zweitens stellen die unterschiedlichen Muster von Symptomkombinationen in den Symptombefunden der Patienten nicht notwendigerweise ein niedrig dimensionales Konstrukt dar, das in nur einer Dimension stabil abbildbar wäre. Falls die Muster an Symptomkombinationen vielfältiger und nur schlecht auf einer Dimension abbildbar sind, dann wird die Abbildung zwangsläufig instabiler und der Einfluss der Fokuspatientenprofile auf die Struktur entsprechend grösser. Zur strukturellen Stabilität der qualitativen Dimension liegen bislang keine Erfahrungswerte vor. Hier wird zukünftige Forschungsarbeit zu leisten sein, bevor die Masse aus der qualitativen Dimension als repräsentativ angesehen werden können.

Ein denkbarer Ausweg aus dem Dilemma der Beeinflussung der Patientenkarte durch die Fokuspatientenbefunde wäre die separate Berechnung von Karten auf der Basis der Referenzpatientenbefunde allein und solche auf der Basis von Referenz- und Fokuspatientenbefunden gemeinsam. Mittels Prokrustes Transformation (z.B. Borg & Groenen, 2005) könnte die gemeinsame Karte auf die ursprüngliche Referenzpatientenkarte zurückgedreht werden (wobei zur Berechnung der Transformationsmatrix ausschliesslich die Referenzpatientenbefunde verwendet würden). Durch ein Anwenden der Transformationsmatrix auf die Koordinaten der Fokuspatientenbefunde könnten diese in den ursprünglichen Raum der Referenzpatienten projiziert werden, ohne deren Positionen zu beeinflussen. Ein derartiges Vorgehen hätte allerdings einen gewichtigen Nachteil: Dadurch, dass die Fokuspatientenbefunde keinen direkten Einfluss auf die Referenzpatientenkarte mehr aufweisen würden, bestünde die Gefahr massgeblich verzerrter Distanzen zwischen den Referenzpatienten- und den Fokuspatientenbefunden. Von besonderer Relevanz für die Praxis wären zu nah an die Fokuspatientenbefunde platzierte Referenzpatientenbefunde, wie im nachfolgenden Absatz eingehend beschrieben wird. Ein „Auseinanderschieben“ von unähnlichen Referenz- und Fokuspatientenbefunden, wie dies in der gemeinsamen NMDS-Lösung vom Algorithmus vorgenommen wird, würde entsprechend nicht mehr stattfinden. Um diese hochgradig praxisrelevanten Fehler zu vermeiden, haben wir uns dafür entschieden, die Patientenkarten als gemeinsame NMDS-Lösung von Referenz- und Fokuspatienten zu berechnen – auch wenn damit eine geringfügige Beeinflussung des Messsystems durch den Messgegenstand in Kauf genommen werden muss.

Anwendung in der klinischen Praxis.

Was einzelne Fokuspatienten betrifft besteht freilich kein Vorbehalt hinsichtlich der Interpretierbarkeit unterschiedlicher Lagen auf der Vertikalen: Die Positionen der Objekte in der Vertikalen werden durch die hauptsächliche qualitative Variabilität der Symptomprofile des Samples bestimmt. Die NMDS findet hierzu den bestmöglichen Kompromiss, die eigentlich mehrdimensionalen qualitativen Unterschiede in den Profilen eindimensional (eben in der Vertikalen) abzubilden. Damit bleibt die hauptsächliche Varianzquelle der Profilunterschiede des verwendeten Samples (Referenz- + Fokuspatientenbefunde) erhalten.

Symptomprofile weit am oberen bzw. unteren Rand der Patientenkarte, weisen generell grössere Distanzen zu den übrigen Befunden auf, als Befunde die in der Mitte der Karte platziert sind. Damit lässt sich auf die Typizität der Symptomprofile rückschliessen: Befunde in der Mitte der Karte weisen offensichtlich typischere Symptomprofile für das vorliegende Sample auf als Befunde am Rand der Karte, denn durch ihre mittige Position werden die Distanzen zu den übrigen Befunden minimiert. Umgekehrt sind Befunde am Rand der Karte i.d.R. untypischer, da deren Distanzen zu den übrigen Befunden maximiert werden. Zusätzlich aber gilt: Befunde am oberen Rand sind zwar untypisch, aber „anders“ untypisch als diejenigen am unteren Rand der Karte.

Für die psychiatrische/psychotherapeutische Praxis dürfte in erster Linie die Ähnlichkeit der Fokuspatientenbefunde zu denjenigen der Referenzpatienten Relevanz besitzen. Wie die bisherige Literatur zeigt, besitzt die konkrete Symptomatik mitunter den stärksten Einfluss auf die Wahl der Antidepressiva in der Behandlung (Zimmerman, 2004). Nun weisen Symptomprofile, die in der Patientenkarte nahe beieinander liegen, auch grosse Ähnlichkeit zueinander auf, sowohl was den Schweregrad als auch was die Qualität der Symptomatik anbelangt. Umliegende Referenzbefunde könnten damit als Indikatoren für die Behandlungsplanung eines Fokuspatienten herangezogen werden. Sofern die Referenzpatienten sowohl mit Ein- als auch mit Austrittsbefund in den Patientenkarten vertreten sind, lässt sich sogar die Effektivität der Behandlungsmassnahmen (falls die entsprechenden Informationen zur Verfügung stehen) ersehen. Eine derart spezifische Anwendung der Patientenkarten, in der die zu einem Fokuspatientenbefund benachbarten Befunde der Referenzpatienten als Informationsgrundlage dienen, bedarf einer detaillierteren Analyse aus NMDS-methodischer Sicht, womit sich der folgende Abschnitt befasst.

Inverse-Interpretation: Ähnlichkeiten von Nachbarobjekten in NMDS-Lösungen

Bei der praktischen Anwendung der Patientenkarten verspricht die lokale Interpretation von Profilähnlichkeiten, d.h. Rückschlüsse von benachbarten Symptomprofilen in der Patientenkarte auf deren Ähnlichkeiten, den grössten Informationsgewinn. Lokale Interpretation wird dabei und im Nachfolgenden als die Beschränkung der interpretierten Distanz-Ähnlichkeitsbeziehungen auf einer kleineren Unterregion des Patientenraumes verstanden (also z.B. die Interpretation der nächsten fünf Nachbarn oder der umliegenden Befunde in einem bestimmten Radius). Denn

so können Informationen zu den Behandlungsmassnahmen bei bereits behandelten (Referenz-)Patienten mit ähnlichem symptomatischem Profil rasch eingesehen und verwendet werden. Allerdings deckt sich die selektive, lokale Interpretation von NMDS Lösungen nicht mehr mit dem globalen Optimierungskriterium der NMDS, auf das hin die Struktur der Befunde modelliert wurde. Das Ziel dieses Abschnitts ist es aufzuzeigen, welche Art von Fehlern durch eine ausschliesslich lokale Interpretation entstehen können sowie, darauf aufbauend, eine Optimierung der Stresswert-Funktion bezüglich lokaler Interpretation zu skizzieren und erste Befunde zu deren Anwendung zu präsentieren. In keiner Weise versteht sich der Abschnitt als abschliessende Evaluation der vorgeschlagenen Optimierung, noch die Optimierung als finale Lösung der Problematik. Ziel ist es vielmehr, eine Richtung aufzuzeigen, in die eine Weiterentwicklung des Algorithmus zur Optimierung von lokalen Strukturen gehen könnte.

Wie im vorangehenden Kapitel erläutert, lassen sich Patientenkarten von unidimensionalen Inventaren derart rotieren, dass die Variabilität in den Symptomprofilen (die qualitativen Unterschiede) entlang der y-Achse abgebildet werden kann. Diese eine Dimension reicht natürlich nicht aus, um die gesamte Variabilität in den Symptomprofilen abzubilden; vielmehr repräsentiert sie die massgeblichsten Unterschiede, welche in den Symptomprofilen des Samples insgesamt vorhanden sind. Eine derartige Reduktion in der Dimensionalität birgt den Vorteil, dass die Hauptvariabilität in den Daten weiterhin bestehen bleibt und damit unabhängiger von zufälligen Unterschieden in den Daten wird. Allerdings entstehen dadurch in der Patientenkarte auch notwendigerweise Strukturvereinfachungen, welche aus klinischer Sicht evtl. nicht gerechtfertigt sind: Sofern sich zwei Profile nicht eben durch diejenigen Symptome unterscheiden, die auch massgeblich zur qualitativen Dimension beitragen, wird deren Distanz in der NMDS Lösung in Bezug auf die eigentlich vorhandene Unähnlichkeit nicht adäquat repräsentiert werden können. Hierin zeigt sich die Hauptproblematik bei lokalen Interpretationen in der Patientenkarte. Während derartige Fehler für die Gesamtstruktur der Patientenkarte minimiert sind, sind sie das bei lokalen Interpretationen keineswegs zwingend. Des Weiteren ist durch die Auswahl der umliegenden Befunde um einen Fokuspatientenbefund ein starker Bias im Fehler zu erwarten. Da die Distanzen zum Fokuspatienten durchwegs gering sind (es werden ja die nächstliegenden ausgewählt) werden die Fehler fast ausnahmslos eine zu hohe Unähnlichkeit im Vergleich zur dargestellten Distanz aufweisen. Es gilt daher bei einer derartigen lokalen Interpretation meist: $d_{ij} < \delta_{ij}$, wobei d die Distanz und δ die Disparität (in der Metrik der Distanzen) bezeichnet. Abbildung 2 zeigt die (normierte) Verteilung der Distanzen und Disparitäten der 15 Fokuspatienten bei insgesamt 72 Befundzeitpunkten zu den 30 Referenzpatienten bei jeweils zwei Befundzeitpunkten des BDI-Samples. Es ist evident, dass zwar viele mittlere Disparitäten vorliegen, die Umsetzung in entsprechende mittlere Distanzen aber nur schlecht gelingt: Die Streubreite der Distanzen (d) bei fixierter Disparität (δ) ist im Bereich [0.7, 1.2] wesentlich grösser als im unteren und oberen Bereich der Disparitätenskala.

Auffällig ist auch die Fehlerverteilung: Der Hauptanteil der Fehler wird durch zu geringe Distanzen im Vergleich zu den Disparitäten verursacht – und zwar bereits bei der globalen Interpretation, also der Interpretation aller Distanzen-Disparitäten-Paare zwischen Fokus- und Referenzpatienten. Zur einfacheren Verständlichkeit werden im Folgenden die Begriffe falsch-positive und falsch-negative Ähnlichkeiten definiert: Falsch-positive Ähnlichkeiten liegen dann vor, wenn eine geringe Distanz eine eigentlich hohe Disparität repräsentiert (Aus einer kleinen Distanz würde fälschlicherweise auf eine hohe Ähnlichkeit geschlossen), falsch-negative Ähnlichkeiten sind entsprechend als grosse Distanzen, die eine geringe Disparität repräsentieren, zu verstehen (Aus einer grossen Distanz würde fälschlicherweise auf eine hohe Unähnlichkeit geschlossen). Bei den normierten Werten in Abbildung 2 repräsentieren Punkte unterhalb der Winkelhalbierenden zwischen x- und y-Achse falsch-positive, während Punkte oberhalb der Winkelhalbierenden falsch-negative Ähnlichkeiten kennzeichnen. Da bei lokaler Interpretation durchgängig von kleinen Distanzen auf die zugrundeliegenden Ähnlichkeiten geschlossen wird, sind falsch-positive Ähnlichkeiten besonders störend (denn falsch negative werden erst gar nicht interpretiert). Diese Problematik wird im Folgenden als Inverses-Interpretationsproblem bezeichnet.

Der hohe Grad an falsch-positiven Ähnlichkeiten bei kleinen und mittleren Disparitäten lässt sich auf die Dimensionsreduktion zurückführen. Um diesen Sachverhalt zu illustrieren, eignet sich ein Gedankenexperiment: Es soll die Distanzmatrix der Standardbasis eines fünf dimensionalen euklidischen Raums per NMDS in einer Dimension abgebildet werden. Das beste (nichttriviale) Resultat erhält man, wenn die Objekte (die Standard-Basisvektoren) mit derselben zwischen-Objektdistanz zueinander und entlang der einen Dimension abgebildet werden. Damit entstehen allerdings, geometrisch bedingt, vier kleine Distanzen, drei mittel-kleine Distanzen, zwei mittel-grosse Distanzen und eine grosse Distanz. Je höher nun die ursprüngliche Dimensionalität der Daten, desto mehr müssen im Prinzip gleichabständige Unähnlichkeiten in kleine Distanzen umgesetzt werden. Mit anderen Worten: Geometrisch bedingt stehen zu wenige mittlere, mittel-grosse und grosse Distanzen zur Rekonstruktion des ursprünglichen Raumes zur Verfügung. Entsprechend weist Abbildung 2 auch besonders viele falsch-positive Ähnlichkeiten bei kleinen und mittel-kleinen Distanzen auf.

Gegensätzlich dazu verhält sich die Verteilung der Profilähnlichkeiten – zumindest unter der Annahme von identisch verteilten Symptomausprägungen. Die grösste Anzahl an Symptombefunden wird eine mittelhohe Unähnlichkeit zu den übrigen Symptombefunden aufweisen: Denn kombinatorisch stehen dafür die meisten Möglichkeiten zur Verfügung. Das in Abbildung 2 ersichtliche Verteilungsmuster zwischen Distanzen und Disparitäten ist also keineswegs als spezifisch für den Datensatz bzw. die Fokuspatienten anzusehen, sondern lässt sich als generelles Prinzip der NMDS festhalten, sofern City-Block Distanzen als Ähnlichkeiten eingesetzt werden.

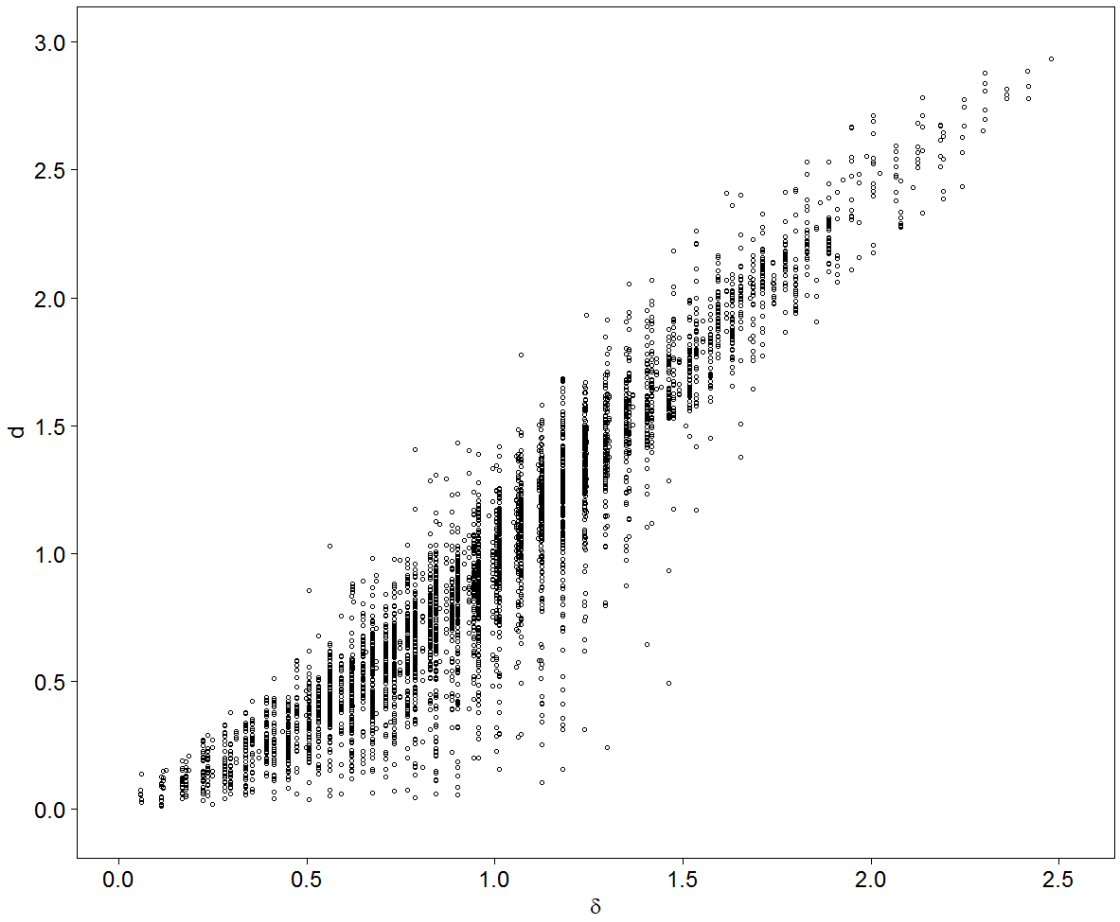

Abbildung 2. Scatterplot der Disparitäten (**δ**) und der Distanzen (*d*) zwischen Fokuspatienten- (*n* = 72) und Referenzpatientenbefunden (*n* = 60).

Spezifika von ausgewählten NMDS Algorithmen in Bezug auf das Inverse-Interpretationsproblem in den Patientenkarten

Die NMDS ist ein Verfahren, das iterativ die bestmögliche Annäherung einer Ähnlichkeitsstruktur an eine geometrische (euklidische) Struktur sucht. Dazu wird eine Loss-Funktion (die Stresswertfunktion) eingesetzt, die zumeist auf das Minimum der quadrierten Abweichungen zwischen den Disparitäten und den zugehörigen Distanzen optimiert wird (Borg, 2005). Im Softwarepaket ProDax (Oberholzer et al., 2008) wird eine davon leicht abweichende Stresswertfunktion eingesetzt (Läge et al., 2005), welche die Abweichungen gewichtet (Gleichung 2).

Die Gewichte w_{ij} werden nicht zu Beginn des Verfahrens festgelegt, sondern während des iterativen Vorgangs der NMDS fortlaufend angepasst. Eine mathematische Beschreibung der Gewichtungsfunktion findet sich in Läge et al. (2005), hierin soll die Beschreibung genügen, dass das Gewicht von grossen Fehlern ab einem bestimmten Schwellenwert rapide abnimmt. Damit werden einzelne (als Messfehler angenommene) grosse Fehler zwischen Disparitäten und Distanzen zugunsten einer besser passenden Gesamtstruktur toleriert.

Das Tolerieren von grossen Fehlern zwischen Disparitäten und Distanzen wird aber vor dem Hintergrund einer lokalen Interpretation zweischneidig. Aus praktischer Sicht sind insbesondere falsch-positive Ähnlichkeiten zu vermeiden, denn diese verleiten zu falschen Rückschlüssen bezüglich Diagnostik und Behandlungsplanung eines Fokuspatienten (Falschinformation). Falsch-negative Ähnlichkeiten sind dagegen weniger gravierend, da daraus lediglich ein Fehlen an Informationszugewinn resultieren kann (denn durch die grosse Distanz zum Fokuspatienten werden diese kaum interpretiert). Aus praktischer Sicht sollte das Minimierungskriterium in den Patientenkarten entsprechend zugunsten einer reduzierten falsch-positiv Rate verschoben werden, wofür sich die Fehlerbehandlung von RobuScal als nicht optimal herausstellt. Selbstredend ist es grundsätzlich denkbar, dass die Fehlerbehandlung von RobuScal, trotz höherer falsch-positiv Rate, zu besseren diagnostischen und behandlungsrelevanteren Informationen führen könnte: Nämlich genau dann, wenn sich diese Informationen ausschliesslich aus der Hauptstruktur des Samples ableiten lassen. Dies als generelles Prinzip zeigen zu können, bleibt allerdings einer zukünftigen Untersuchung vorbehalten und wird im vorliegenden Manuskript nicht weiter Thema sein.

Borg und Groenen (2005) stellen unterschiedliche Stresswertfunktionen vor, von denen eine logarithmische Variante auf den ersten Blick für das vorliegende Problem adäquat erscheint. Statt die Minimierung der quadrierten Fehler anzustreben, wird dabei die Minimierung der quadrierten Fehler des Logarithmus der Disparitäten und der Distanzen angezielt. Damit werden Abweichungen bei kleinen Disparitäten stärker „gewichtet“ als Abweichungen bei grossen Disparitäten. Allerdings zeigt sich auf den zweiten Blick, dass zwar kleine Disparitäten besser in kleine Distanzen umgesetzt werden, aber für grosse Disparitäten, die durch kleine Distanzen repräsentiert werden, weiterhin kein Unterschied in der Gewichtung besteht. Die Hauptproblematik der falsch-positiven Ähnlichkeiten wird entsprechend nur ungenügend abgeschwächt.

Die Gewichtung in der Stresswertfunktion, die im Folgenden vorgeschlagen wird, bezieht sich daher nicht länger auf die Disparitäten, sondern direkt auf die darzustellenden Distanzen (Distanzengewichtung). Sie folgt der Stresswertfunktion von RobuScal (Gleichung 2), bestimmt allerdings die Gewichte nicht aufgrund des Fehlers, sondern aufgrund der darzustellenden Distanzen. Die Gewichte könnten z.B. gegeben werden durch:

$$w_{ij}(d'_{ij}) := \begin{cases} \frac{1}{\mathrm{Med}\left(d'_{ij}\right)\cdot 0.25^2}, & \frac{d'_{ij}}{\mathrm{Med}(d'_{ij})} < 0.25 \\ \frac{\mathrm{Med}\left(d'_{ij}\right)}{d'^2_{ij}}, & \frac{d'_{ij}}{\mathrm{Med}(d'_{ij})} \geq 0.25 \end{cases} \tag{3}$$

damit ist w_{ij} immerhin stetig und definiert auf ganz $\mathbb{R}^+$, solange gilt, dass $\mathrm{Med}(d'_{ij}) > 0$. Die Bestimmung der Distanzen d' erfolgt, analog zur Gewichtung in RobuScal, in jedem fünften Iterationsschritt und bleibt dazwischen unverändert (Läge et al., 2005).[6]

Praktisch relevant ist eine festgelegte maximale Gewichtung (ein Viertel der Mediandistanz), welche einen übermässigen Einfluss von Einzeldistanzen verhindert und die Gewichtungsfunktion auch bei Distanzen mit Wert 0 definiert. Von einem inhaltlichen Standpunkt aus betrachtet, wird damit kleinen und kleinsten Distanzen dieselbe Relevanz zugeschrieben. Es wird also quasi ein Interpretationsradius um die Objekte festgelegt (der einem Viertel der Mediandistanz entspricht), innerhalb dessen den Objekten dasselbe, maximale Gewicht zugeordnet wird. Das Fehlen einer äquivalenten Obergrenze entspringt ebenfalls inhaltlichen Überlegungen: Grosse Distanzen in der Struktur sind zunehmend irrelevanter für die lokale Interpretation. Für eine Begrenzung des minimalen Gewichts ist entsprechend keine ähnliche Inhaltliche Rechtfertigung gegeben.

Im Vergleich zeigt sich ein deutlicher Effekt: Die Streubreite der Disparitäten (δ) bei kleinen Distanzen (d) wird merklich reduziert. Die aus praktischer Sicht besonders schwerwiegenden falsch-positiven Ähnlichkeiten, also kleiner Distanzen, mit grossen Disparitäten der Profile, können mit der distanzgewichteten Stresswertfunktion (Gleichung 3) substantiell reduziert werden (Abbildung 3).

Neben der algorithmischen Optimierung hin zu einer valideren lokalen Interpretation drängt sich die Frage auf, ob nicht u.U. die Dimensionalität der Lösung angepasst werden sollte. Denn eine Erhöhung der Dimensionalität führt generell zu werttreueren Abbildungen der Ausgangsstruktur. Im folgenden Abschnitt werden daher drei unterschiedliche Stresswertfunktionen sowohl im 2-dimensionalen als auch im 3-dimensionalen Raum verglichen, um erste Anhaltspunkte zur Adäquanz der Abbildungen zu erhalten.

[6] Die vorgestellte Gewichtungsfunktion (Gleichung 3) ist eher als generelles Prinzip zu verstehen denn als finale Funktion. Insbesondere die Begrenzung der maximalen Gewichte, aber unter Umständen auch der Exponent im Nenner der Funktion, sollte empirisch überprüft und den Resultaten entsprechen angepasst werden.

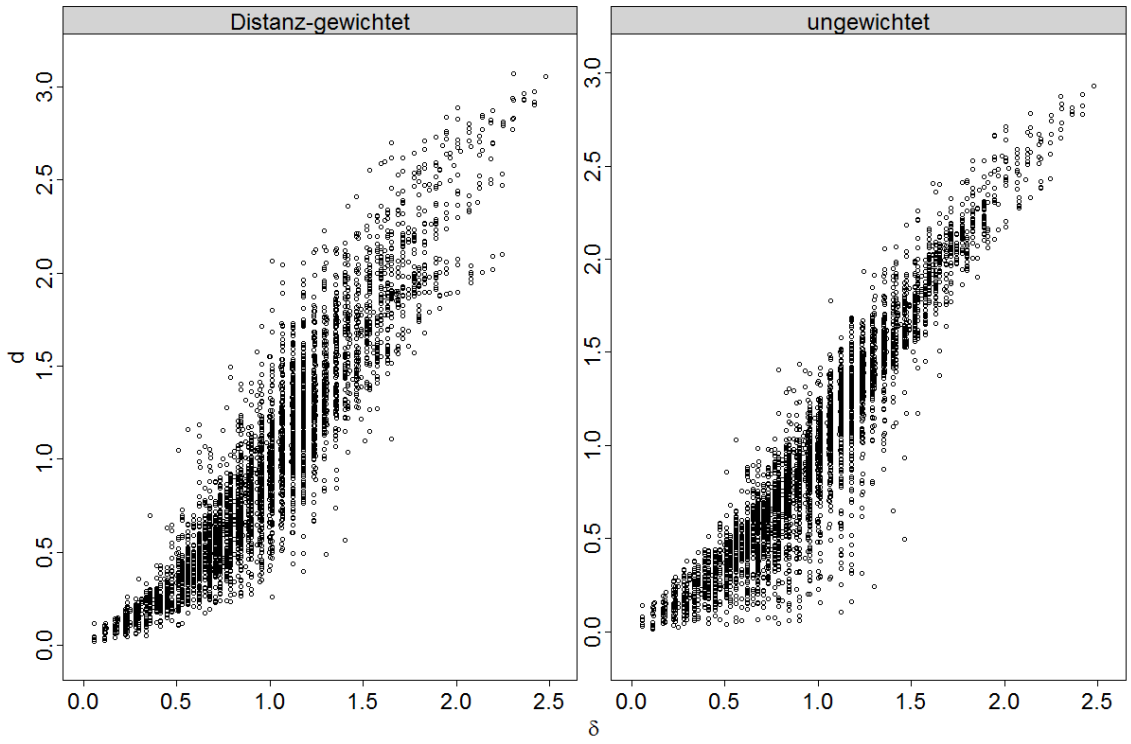

Abbildung 3. Scatterplot der Disparitäten (**δ**) und der Distanzen (d) zwischen Fokuspatienten- (*n* = 72) und Referenzpatientenbefunden (*n* = 60) bei unterschiedlichen Stresswertfunktionen. Links wurde die in Gleichung 3 dargestellte, rechts eine ungewichtete Stresswertfunktion verwendet.

Der Einfluss der Stresswert-Funktion und der Dimensionalität auf das Inverse-Interpretationsproblem in Patientenkarten

Ein algorithmischer Ansatz zur Lösung des Inversen-Interpretationsproblems, wie im vorhergehenden Abschnitt beschrieben, stellt selbstverständlich nicht den einzigen Lösungsansatz dar. Ebenso könnte auch durch eine Erhöhung der Dimensionalität eine bessere Abbildung der Datenstruktur erreicht werden. Durch die Erhöhung der Dimensionalität wird allerdings auch die Darstellung des NMDS Raumes immer schwieriger. Für einen angezielten praktischen Einsatz hätte dies schwerwiegende Konsequenzen: Während die Darstellung von Patientenkarten im 3-dimensionalen Raum zwar noch als ein Raum möglich wäre, gelänge dies im 4- oder gar 5-dimensionalen Raum nicht mehr. Die Struktur könnte zwar durch die sukzessive Projektion der Objekte auf orthogonale Ebenen des Raumes (sogenannte Unterräume) weiterhin abgebildet werden, die Interpretation der Struktur über die verschiedenen Abbildungen hinweg würde damit aber um ein vielfaches komplexer. Die folgende Analyse hatte zum Ziel, erste Abschätzungen zur Eignung der unterschiedlichen Stresswertfunktionen sowie zur Relevanz der Dimensionalität bezüglich des Inversen-Interpretationsproblems zu erhalten. Um die Interpretierbarkeit

der Patientenkarten nicht zu vernachlässigen, wurden nur 2- und 3-dimensionale Räume untersucht.

Sample

Als sample wurde der bereits in der Einleitung vorgestellte BDI-Datensatz verwendet.

Vorgehen

Aus den 45 Patienten wurden zufällig 30 Patienten gezogen, deren Ein- und Austrittsbefunde als Referenzpatientenbefunde dienten. Für die übrigen 15 Patienten wurden jeweils individuelle Patientenkarten berechnet, welche aus den 60 Referenzbefunden, plus den wöchentlichen Befunden des jeweiligen Fokuspatienten (variierende Anzahl je nach Patient), plus einem 0-Befund (Nullsymptomatik) berechnet wurden. Der 0-Befund war durch die Absenz aller symptomatischen Ausprägungen charakterisiert (lauter nullen) und diente als „Interpretationsanker“ in der Karte. Als Proximitätsmass für die Ähnlichkeitsbeziehungen zwischen den Befunden dienten City-Block Distanzen.

Die Berechnung der Patientenkarten erfolgte unter Variation der beiden Faktoren Stresswertfunktion und Dimensionalität. Es wurden drei verschiedene Stresswertfunktionen (RobuScal, Ungewichtet, Distanz-gewichtet) und zwei verschiedenen Dimensionalitäten (2-dimensional, 3-dimensional) evaluiert. Entsprechend wurden von allen 15 Proximitätsmatrizen (der 15 Fokuspatienten) jeweils sechs NMDS Lösungen berechnet. Für alle Varianten wurde jeweils die Distanzmatrix extrahiert, welche anschliessend als Grundlage zur Bestimmung der Nachbarobjekte der Fokuspatientenbefunde diente.

Zur nachträglichen Bestimmung der Unähnlichkeit zwischen zwei Befunden wurde eine mittelwertszentrierte, normierte City-Block-Distanz (CB_z) verwendet. Dies deshalb, weil die Abbildbarkeit des Schweregrads in der Karte ausser Zweifel steht (wie im Absatz zum Verlauf in den Patientenkarten berichtet beträgt die Korrelation zwischen x-Achse und Schweregrad praktisch 1 (r=0.988)) und um Konfundierung der Profilähnlichkeit mit dem Schweregrad zu vermeiden. Das Mass CB_z wurde definiert als:

$$CB_z = \frac{1}{d}\sum_{i=1}^{d}|(x_i - \bar{x}) - (y_i - \bar{y})| \quad (4)$$

Wobei *x* und *y* zwei BDI-Befunde und *d* die Dimensionalität des Fragebogens repräsentieren; im vorliegenden Fall also *d* = 21. Da mit zunehmendem Schweregrad generell auch die Profilunähnlichkeit nach dem Mass CB_z zunimmt, wurden die Fokuspatientenbefunde in drei Schweregradkategorien unterteilt, für welche die Resultate jeweils gesondert berichtet werden. Die Schweregradkategorien wurden anhand des Gesamtscores im BDI in drei Gruppen unterteilt: gering [0,11), mittel [11,18) und hoch [18,63]. Die Schweregradkategorien folgten damit

der Einteilung zur klinischen Relevanz von Hautzinger, Bailer, Worall und Keller (1995) wonach die tiefe Kategorie klinisch unauffällig, die mittlere auf leichte bis mässige Symptomatik und die schwere Kategorie auf klinische Relevanz der Symptome hindeutet. In Tabelle 1 sind die Verteilung der Anzahl an Fokuspatientenbefunden je Schweregradkategorie sowie minimale, durchschnittliche und maximale Werte des Unähnlichkeitsmasses CB_z zusammengefasst.

Die CB_z Masse je Stresswertfunktion und Dimensionalität wurden in Bezug auf zwei unterschiedliche Kriterien ausgewertet. Zum einen wurde eine bestimmte Anzahl an Nachbarschaftsbefunden (nächstliegende *x* Referenzbefunde in der NMDS, unabhängig von deren Distanz) auf ihre Ähnlichkeit zum Fokuspatientenbefund hin überprüft. Zum anderen wurden Maximaldistanzen definiert (unabhängig von der Anzahl der Befunde innerhalb des Radius‘), innerhalb derer alle Referenzbefunde auf ihre Ähnlichkeit zum Fokuspatientenbefund überprüft wurden. Die beiden Interpretationsansätze widerspiegeln unterschiedliche lokale Interpretationsstrategien in der Karte, die beide sowohl als Rein-, am häufigsten aber sicherlich als Mischform in der praktischen Anwendung auftreten dürften.

Tabelle 1

Kennwerte zur Verteilung der Fokuspatienten-Befunde und deren Unähnlichkeitskoeffizienten ($\boldsymbol{CB_z}$) in den drei Schweregradkategorien gering, mittel und hoch

Schweregradkategorie	Anzahl Befunde je Kategorie (*N* = 72)	Min(CB_z)	$\overline{CB_z}$	Max(CB_z)
hoch	26	0.46	0.78	1.29
mittel	18	0.35	0.73	1.18
gering	28	0.17	0.64	1.17

Resultate

Abbildung 4 zeigt die mittlere Profilähnlichkeit (CB_z) zwischen den Fokuspatientenbefunden und den jeweils dazu am nächsten liegenden fünf Referenzpatientenbefunden. Die Distanz zu diesen fünf Nachbarn spielte dabei keine Rolle. Die Fokuspatientenbefunde wiesen in den 3-dimensionalen NMDS-Lösungen generell etwas geringere Unähnlichkeitswerte zu Ihren Nachbarbefunden auf als in den 2-dimensionalen Lösungen (Abbildung 4). Die Stresswertfunktionen unterschieden sich ebenfalls systematisch (Tabelle 2) hinsichtlich der mittleren Unähnlichkeit der Nachbarbefunde, wobei die distanzgewichtete Stresswertfunktion generell die niedrigsten Unähnlichkeitswerte zeigte (Abbildung 4).

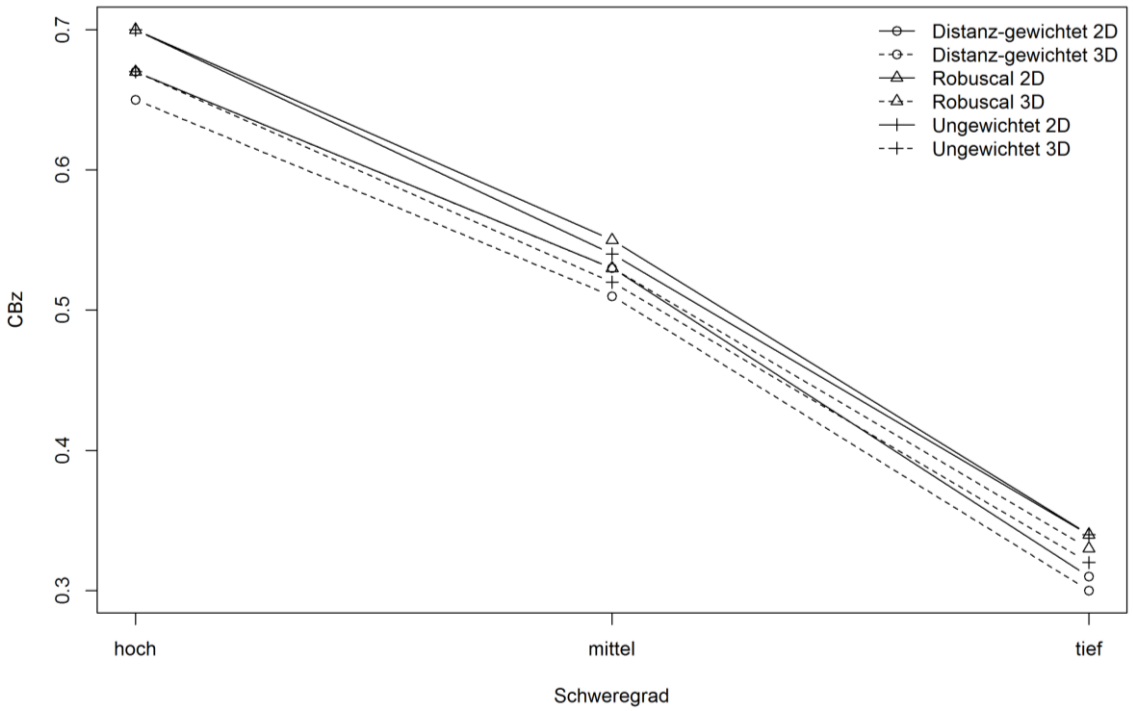

Abbildung 4. Mittelwert der Unähnlichkeiten ($\boldsymbol{CB_z}$) der jeweils nächsten fünf Nachbarn zu jedem Fokuspatientenbefund (N = 72), unabhängig von der effektiven Distanz zwischen Fokuspatienten- und Referenzpatientenbefund.

In einer zweiten Analyse wurden nur Unähnlichkeitsmasse zu denjenigen Referenzpatientenbefunden einbezogen, die in einem bestimmten, maximalen Radius zum jeweiligen Fokuspatientenbefund positioniert waren. Die Festlegung der Maximaldistanz musste willkürlich vorgenommen werden, da noch keinerlei Erfahrungswerte zu lokalen Interpretationen in Patientenkarten vorliegen. Es wurde allerdings darauf geachtet, dass jede Zelle der Stresswert-Schweregradkombinationen mindestens 10 Ausprägungen aufwies, um einen einigermassen stabilen Mittelwert zu erhalten. Die maximale Interpretationsdistanz wurde auf d = 0.4 festgesetzt.[7]

Im Gegensatz zu den Unähnlichkeitswerten der umliegenden fünf Nachbarn, zeigten die Werte in Abhängigkeit einer Maximaldistanz ein weitaus differenzierteres Bild (Abbildung 5). Es zeigte sich ein substantieller Einfluss der Dimensionalität, insbesondere bei hohem Schweregrad. Der Einfluss der verwendeten Stresswertfunktion nahm ebenfalls deutlich zu. Die Variabilität im Unähnlichkeitswert, besonders bei hohem Schweregrad, fiel deutlich höher aus als beim Einbezug einer fixen Anzahl an Nachbarn (Abbildung 4).

[7] Zur Klärung: die Distanzen in der NMDS sind dermassen normiert, dass die mittlere Distanz der Objekte zum Schwerpunkt der Karte genau 1 betragen.

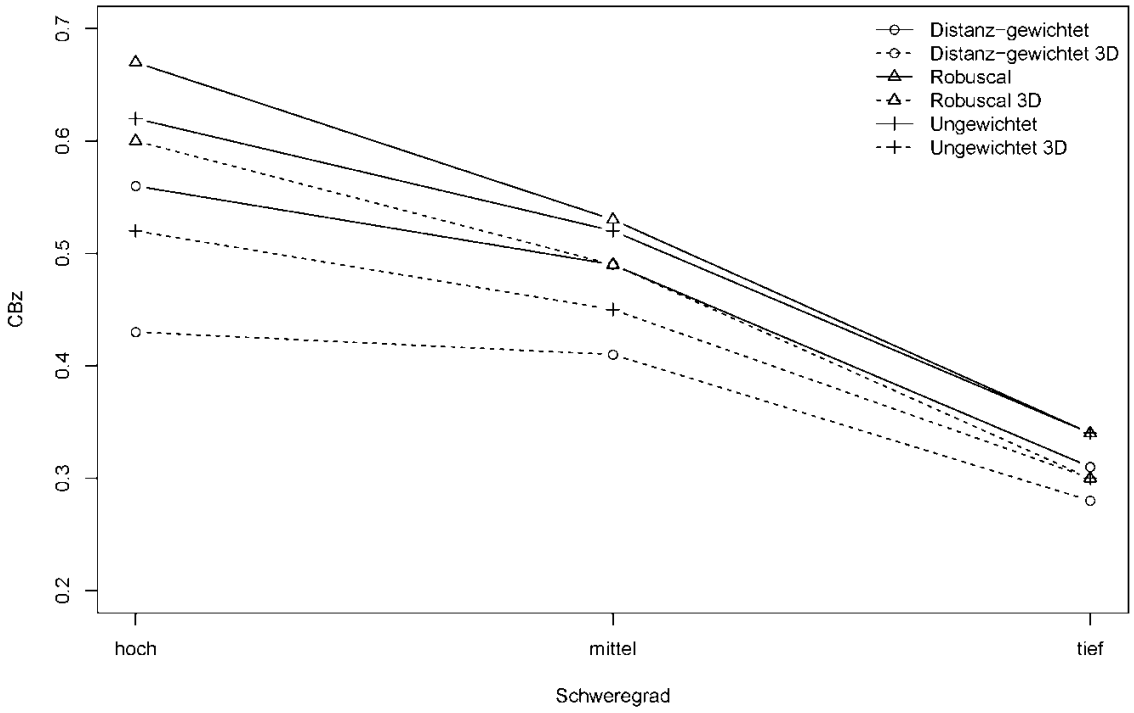

Abbildung 5. Mittelwert der Unähnlichkeit ($\boldsymbol{CB_z}$) zwischen Fokuspatientenbefund (N = 72) und Referenzpatientenbefunden. Als Kriterium für den Einbezug der Referenzpatientenbefunde diente eine maximale Interpretationsdistanz (d = 0.4) in der Patientenkarte.

Kurzdiskussion der Resultate

Die Resultate in Abbildung 4 und 5 zeigten systematische Unterschiede bei der Verwendung unterschiedlicher Stresswertfunktionen und Dimensionalitäten auf. Lokale Interpretationen, also der Rückschluss von kleinen Distanzen in den Patientenkarten auf kleine Disparitäten zwischen den Objekten, scheinen durch die Verwendung einer distanzgewichteten Stresswertfunktion valider zu werden, was die Resultate in den Abbildungen 4 und 5 nahelegen. Besonders deutlich fielen die Unterschiede zwischen den Stresswertfunktionen unter Verwendung einer maximalen Interpretationsdistanz (Abbildung 5) aus. Die Erhöhung der Dimensionalität von zwei auf drei zeigte, wie erwartet, ebenfalls bessere Resultate in den mittleren Profilunähnlichkeiten. Auch im 3-dimensionalen Patientenraum wies die distanzgewichtete Stresswertfunktion, im Vergleich zur ungewichteten Stresswertfunktion und zu RobuScal, die kleinsten Unähnlichkeitswerte auf. Eine Erhöhung der Dimensionalität ist allerdings mit erheblichen Schwierigkeiten bei der Darstellung und der Interpretation der Distanzen verbunden: Eine Empfehlung bezüglich der zu verwendenden Dimensionalität kann daher in diesem Artikel nicht ausgesprochen werden. Des Weiteren scheint es plausibel, dass sich die Notwendigkeit weiterer Dimensionen von Inventar zu Inventar unterscheidet. Es wird erwartet, dass der BDI (welcher den Daten dieser Studie zugrunde liegt), als unidimensionales Inventar, mit einer geringeren Zahl an Dimensionen abgebildet werden kann, als psychopathologisch breitere, multidimensionale Inventare.

Diskussion

Die selektive Interpretation von Distanzen in Bezug auf deren repräsentierte Ähnlichkeiten in NMDS-Lösungen unterliegt systematischen Verzerrungen, wie die vorherigen Abschnitte zum Inversen-Interpretationsproblem aufzeigen konnten. Von primärem praktischem Interesse werden kleine Distanzen in den Patientenkarten sein, um so die ähnlichsten Referenzfälle zu identifizieren. Besonders relevant für die Anwendung der Patientenkarten in der Praxis sind deshalb falsch-positive Ähnlichkeiten, denn diese verleiten fälschlicherweise zum Rückschluss von kleinen Distanzen auf hohe Profilähnlichkeiten. Derart lokale Interpretationen von Objektdistanzen werden allerdings von den gängigen Stresswertfunktionen, welche die Patientenkarten durchwegs bezüglich der Gesamtstruktur der Daten optimieren, nicht berücksichtigt. Die in diesem Artikel vorgeschlagene distanzgewichtete Stresswertfunktion beschreibt einen Ansatz, wie die lokalen Zusammenhänge zwischen den Objekten stärker in die Modellierung einbezogen werden können. Die Analyse zur Eignung der vorgeschlagenen distanzgewichteten Stresswertfunktion zeigte, dass durch die Gewichtung lokal validere Strukturen erzeugt werden als dies RobuScal, mit der robusten Stresswertfunktion (Gleichung 2), oder eine ungewichtete Stresswertfunktion vermögen.

Einschränkend zu den Resultaten im letzten Abschnitt muss allerdings erwähnt werden, dass die vorliegenden Auswertungen auf einem rein technischen Ähnlichkeitsmass beruhten, welches nicht zwingend mit der klinisch relevanten Ähnlichkeit übereinstimmen muss. Es ist entsprechend denkbar, dass sich für den Einsatz der Patientenkarten in der klinischen Praxis die Stresswertfunktion von RobuScal oder gar eine ungewichtete Stresswertfunktion besser eignen könnten – trotz der gegenteiligen Befunde, unter Verwendung des technischen Ähnlichkeitsmasses.

Das Inverse-Interpretationsproblem tritt nicht allein bei lokaler Interpretation von NMDS-Lösungen auf: Durch die Reduktion der Dimensionalität resultiert allgemein eine hohe, geometrisch bedingte, falsch-positiv Rate. Auch wenn extreme falsch-positive Fehler durch den Einsatz einer distanzgewichteten Stresswertfunktion vermieden werden können, so werden zukünftige Studien weisen müssen, ob überhaupt alle behandlungsrelevanten Charakteristika von Symptombefunden in niedrig dimensionalen Räumen abgebildet werden können.

Immerhin zeigen sich durchwegs substantiell ähnlichere Befunde in der unmittelbaren Umgebung (Abbildung 5) als im Mittel aller Befunde (Tabelle 1). Dies spricht generell für die Tauglichkeit der Patientenkarten zur Separierung von systematisch unterschiedlichen Symptombefunden. Eine Anwendung der Patientenkarten in der klinischen Praxis scheint vor diesem Hintergrund vielversprechend. Schliesslich fand Zimmerman (2004), dass in erster Linie symptomatische Information für die Auswahl der Psychopharmaka in psychiatrischen Behandlungen herangezogen wird: Durch ähnliche Symptomprofile zweier Patienten, eines bereits behandelten

Referenzpatienten und eines noch unbehandelten Fokuspatienten, wären so z.B. Rückschlüsse vom verwendeten Psychopharmaka des Referenzpatienten zum Einsatz beim Fokuspatienten denkbar. Sollten die Referenzpatienten, ähnlich der vorliegenden Analyse, sowohl mit Ein- als auch mit Austrittsbefund in der Patientenkarte vorliegen, wären darüber hinaus sogar Abschätzungen bezüglich der Effektivität der Behandlung möglich. Die Positionen der Referenzpatienten vor und nach der Behandlung können dazu als Indikatoren für den Erfolg der Behandlung herangezogen werden: Bei standardisiert rotierten Patientenkarten liesse sich die Reduktion im Schweregrad als zurückgelegte Distanz auf der x-Achse der Patientenkarte ersehen (Abbildung 1). Ein Einsatz in der Qualitätskontrolle, aber auch zur generellen Förderung der Evidence-Based Medicine in der Psychiatrie scheint vor diesem Hintergrund vielversprechend.

Bezüglich klinischer Anwendbarkeit stellt die Standardisierung der Patientenkarten, insbesondere während einer Initialphase, vermutlich eine substantielle Hilfestellung dar. Denn die Invarianz der NMDS-Lösungen gegenüber Skalierung, Rotation, Translation und Spiegelung, also der Umstand, dass die Semantik der Struktur ausschliesslich in den Distanzrelationen der Objekte zu finden ist, birgt interpretative Hürden. Da der Schweregrad bei eindimensionalen Inventaren eine stabile Dimension definiert, könnte die Ausrichtung (z.B. der x-Achse) der Patientenkarten an dieser Dimension als Standard fungieren. Die weiteren Dimensionen (je nach Anzahl der verwendeten Dimensionalität in den Patientenräumen) würden dann in erster Linie durch die Profile der Symptombefunde definiert. Durch die in Robuscal verwendete Startkonfiguration ist eine grobe Ausrichtung des Schweregrads entlang der x-Achse bereits gegeben: Für die Startkonfiguration der NMDS wird eine Hauptkomponentenanalyse für die Platzierung der Objekte verwendet und sofern City-Block Distanzen und ein breiter Bereich an Schweregraden als Basis für die NMDS dienen, definiert der Schweregrad eine stabile Hauptkomponente. Diese Schweregraddimension ist sogar derart stabil, dass sie den Patientenraum auch für unterschiedliche Datenbasen bezüglich Schweregrad weitestgehend Skaleninvariant macht. Ein fixer „Ankerpunkt“ wie die Nullsymptomatik könnte zudem als Rotationskriterium herangezogen werden, sodass eine Reduktion in der Symptomatik z.B. immer als Bewegung von links nach rechts dargestellt werden kann.

Weiter wurde argumentiert, dass die Patientenkarten von unidimensionalen Inventaren stets eine charakteristische Objektverteilung aufweisen: Eine kometenhafte Form mit stark geclusterten Austrittsbefunden (Komet) und breit gestreuten Eintrittsbefunden (Schweif). Dies erleichtert die einfache visuelle Interpretation in den Patientenkarten auch ohne Standardisierung der Achsen, setzt aber zwingend eine grosse Breite an unterschiedlichen Schweregraden im Referenzsample voraus. Im hierin verwendeten Referenzsample wurde dieser hohen Variabilität im Schweregrad durch die Verwendung von Ein- und Austrittsbefunden Rechnung getragen.

Der vorliegende Artikel weist durch seine explorative Ausrichtung eine grosse Zahl an Einschränkungen auf, auf die an dieser Stelle noch einmal gesammelt hingewiesen werden soll. Erstens wurden die Patientenkarten im vorliegenden Artikel auf die Datenbasis von unidimensionalen Inventaren beschränkt. Auswirkungen dieser Einschränkung sind sicherlich bei der Objektverteilung (Kometenform) in den Patientenkarten zu erwarten. Es ist anzunehmen, dass Patientenkarten auf der Basis von multidimensionalen Inventaren wesentlich breiter gestreute Symptomprofile aufweisen. Weiter ist anzunehmen, dass die Erhöhung der Dimensionalität der NMDS bei multidimensionalen Inventaren dringlicher sein dürfte als bei unidimensionalen Inventaren. Denn die Variabilität der Symptomprofile wird bei allgemeinen psychopathologischen Inventaren (wie z.B. dem AMDP) weitaus grösser sein, was eine Abbildung in niedrig dimensionalen Räumen zusätzlich erschwert. Ausserdem wurde in der vorliegenden Analyse die Anzahl an Referenzbefunden konstant gehalten (in der Analyse wurden insgesamt 60 Referenzbefunde verwendet). Die Grösse und Repräsentativität des Samples wird massgeblich zur Stabilität der qualitativen Dimensionen beitragen, was in der vorliegenden Analyse nicht überprüft wurde. Dies in erster Linie deshalb, weil die Überprüfung der Stabilität ein separates Versuchsdesign erfordert hätte, was den Umfang des vorliegenden Manuskripts vollends gesprengt hätte. Der Vergleich von Fokuspatientenbefunden mittels Kennwerten der qualitativen Dimension kann entsprechend, ohne vorherige Validierung der Stabilität bzw. dem Einsatz robuster Methoden noch nicht empfohlen werden.

Trotz den genannten Einschränkungen sind wir überzeugt, dass die Patientenkarten einen substantiellen Informationsgewinn für die praktische Behandlung darstellen. So ist es denkbar, dass das alleinige, einfach interpretierbare Feedback zum Verlaufs eines Fokuspatienten bereits erheblichen Einfluss auf die Behandlungsqualität und -effektivität ausübt (Hannan et al., 2005; Slade et al., 2008). Die bisherige Auswertetradition der psychopathologischen Inventare zieht für diesen Zweck leider die qualitativen Unterschiede (Symptomprofile) zwischen den Befunden nur mangelhaft mit ein. So existiert z.B. im HAMD (Hamilton, 1960) oder dem BDI (Beck et al., 1961) nur ein einziger Summenwert, der den Gesamtschweregrad der Depression repräsentiert; dies trotz einer Vielzahl an faktorenanalytischen Befunden, die den Inventaren zusätzliche systematische Varianz bescheinigen (z.B. Bagby, Ryder, Schuller, & Marshall, 2004; Bühler, Seemüller, & Läge, 2013; Ward, 2006; Bühler, Keller, & Läge, 2012) und trotz der weitverbreiteten auch theoretischen Kritik zur fehlenden Homogenität der Depression (z.B. Shorter, 2007; Damm, Eser, Schüle, Möller, Rupprecht, & Baghai, 2009). Im AMDP (Arbeitsgemeinschaft für Methodik und Dokumentation in der Psychiatrie, 1981) sind die Symptome in mehrere Gruppen (die Syndrome) geordnet, wobei auch diese Syndrome die symptomatischen Unterschiede zwischen den Befunden nur unzulänglich abbilden. Denn jedes Syndrom im AMDP besteht im Mittel wieder aus rund 8 Symptomen, von denen nur der Summenwert bekannt ist. Dass diese systematischen Unterschiede im Symptomprofil für die Statusdiagnostik

genutzt werden können, haben Egli et al. (2009) anhand von Patientenkarten sehr schön aufzeigen können. Die hier präsentierte Auseinandersetzung mit den Patientenkarten, mit besonderem Blick auf unidimensionale Skalen, setzt die Arbeit von Egli et al. (2009) fort, indem sie zusätzliches Potential der Patientenkarten in der Verlaufsdiagnostik aufzeigt. Die bestehenden Studien zur Wirkung von Feedback in der psychotherapeutischen Behandlung (Slade et al., 2008) und zur Auswahl von Psychopharmaka (Zimmerman, 2004) lassen die Hoffnung zu, dass NMDS Patientenkarten künftig einen substantiellen, positiven Einfluss auf die Effektivität von psychiatrischen und psychotherapeutischen Behandlungen haben werden.

Conclusion

The studies included in this thesis added new findings to the body of evidence in depression research as well as in NMDS methodology. The diverging findings of the BDI-II's symptom structure in the literature (e.g. Brouwer et al., 2013; Quilty et al., 2010) could be explained well by the NMDS results of the study entitled *"Die Symptomstruktur des BDI-II: Kernsymptome und qualitative Facetten"*, which clearly indicated a factorial complex symptom structure. Moreover, the results suggested an additional, activation related factor, which was replicated in a second study entitled *„The Symptom Structure of the Hamilton Depression Scale (HAM-D): an NMDS analysis"*. Thus, differences seem to exist in the associated activation levels of depressive symptoms. These findings endorse the historical dichotomy of agitated and retarded depression, which has been postulated time and again (Koukopoulos & Koukopoulos, 1999; Shorter, 2007). The specific benefits of NMDS in the analysis of symptom structures were illustrated in all three studies at the beginning of this thesis. The capabilities of NMDS to dimensionally model similarity data with very little assumptions regarding the data's distribution makes it an excellent tool to review factor analytic results. Two different approaches were applied in this thesis. Firstly, NMDS analyses were based on the similarities between the symptoms in a sample of depressive patients: the distribution of the symptoms in the NMDS solution (throughout this thesis two dimensional NMDS solutions were applied) directly reflected the underlying factor structure. Secondly, co-occurrences of symptoms associated with the same factors in the literature were applied as similarity data: this procedure allowed replicating the structure through a meta-analytic procedure described by (Frick et al., 1993; Loeber & Schmaling, 1985).

The findings from "Die Symptomstruktur des BDI-II: Kernsymptome und qualitative Facetten" and "The Symptom Structure of the Hamilton Depression Scale (HAM-D): an NMDS analysis" have contributed to the suggestion that depression may comprise qualitatively different conditions – a hypothesis which has been heavily debated in the past few decades (e.g. Baumeister & Parker, 2012; Fava et al., 1997; Lichtenberg & Belmaker, 2010). The concern of different conditions in depression primarily relates to the high rate of treatment non-responders and high inter-individual differences in the efficacy of treatment (Cuijpers, Straten, Bohlmeijer, Hollon, Andersson, & van Straten, 2010; Turner et al., 2008). Many authors have hoped to disentangle these conditions and thus to be able to define specific treatment regimens for the patients' individual conditions (e.g. Baumeister & Parker, 2012; Carragher et al., 2009; Lichtenberg & Belmaker, 2010), however, with limited success so far (Esposito & Goodnick, 2003; Fava et al., 1997).

The results of the study entitled "The predictive power of subgroups: an empirical approach to identify depressive symptom patterns that predict response to treatment" indicated

five qualitatively different symptom patterns in a large sample of HAM-D data. The current study analyzed previously centered data, which allowed disentangling the severity and the quality of depression. In contrast, previous studies applied LCA to non-centered data (Carragher et al., 2009; Chen et al., 2000; Eaton et al., 1989; Kendler et al., 1996; Sullivan et al., 2002), which led to strongly pronounced severity differences between the derived classes. Subsequently, the current study applied an LME analysis, which included the grouping variable from the LCA as a predictor for response to treatment. Patients associated to one specific class revealed significantly slower responses to treatment. These results suggested that specific symptom patterns indeed yield predictive power with respect to treatment response.

The results are promising for depression research as well as for the application of NMDS in the analysis of individual patient profiles. Although many studies tried to statistically delineate different conditions (e.g. Aggen et al., 2005; Blazer, 1989; Carragher et al., 2009; Cox et al., 2001;), their impact on the concept of depression has been small and their efforts have even been considered as unsuccessful by some authors (e.g. Lichtenberg & Belmaker, 2010). However, the results of the study included in this thesis not only proposed a new classification of depressive patients, but they also suggested considerably slower response rates to treatment for one of the classes. Said class yielded a symptom pattern similar to melancholic depression.

Hence, different symptom patterns were identified, which may serve as promising criteria to delineate different depressive conditions. Furthermore, the findings pave the way for an application of NMDS analyses of individual patients' symptom profiles. After all, the main advantage of patient maps follows the premise that the quality of a patients' disorder (i.e. the specific symptom pattern) adds supplemental, treatment relevant information to the severity of the disorder. However, a major limitation of the study was that only prognostic predictions (i.e. predictions of response to a particular treatment) but no prescriptive predictions (i.e. predictions of differential response to one versus the other treatment) could be obtained. Nevertheless, the prognostic differences in treatment response yield great potential that another treatment modality may indeed be more successful. Hopefully, patient maps will be able to separate these different groups of patients and will therefore be able to grant the clinician access to treatment relevant information about a patients' condition.

However, there are many considerations in the analysis of individual symptom profiles within a patient maps framework. One limitation, i.e. a false-positive bias in the inference of profile similarity, was extensively discussed in the last manuscript entitled *"Berechnung und Interpretation von NMDS Patientenkarten für die Verlaufsdiagnostik: Erste Befunde"*. The proposed adoption of a distance weighted stress function showed good results in the conducted analysis, however, some preferable mathematical properties of the function were not explored (e.g. proof of convergence). Thus, the manuscript rather proposes a first approach than a final solution for the false-positive inference bias. Despite the methodological hurdles discussed in

the manuscript, an application of NMDS in clinical practice with respect to continuous feedback during the course of treatment seems promising: first results suggested reasonably homogeneous profiles in the near of a focus-patient's location.

Looking beyond the mere application of NMDS to depressive inventories, the manuscripts included in this thesis pushed the boundaries of NMDS methodology and methods to analyze item structures of questionnaires. The manuscript entitled "*Better bootstrap NMDS analyses – confidence regions and improved location estimates in Nonmetric Multidimensional Scaling"* confirmed the applicability of bootstrap procedures in NMDS analyses. By applying bootstrap procedures, instable item structures can be identified by large confidence regions in the respective NMDS solutions. Furthermore, non-overlapping confidence regions suggest statistically interpretable differences in the locations of the items, enhancing the validity of the interpretation of a structure. Additionally, an extended procedure based on the bootstrap has been shown to reduce the bias in the calculation of the similarity data, which generally increases the validity of the method.

Furthermore, the manuscript *"Activation as an overlooked factor in the BDI-II: a factor model based on core symptoms and qualitative aspects of depression"* examined the structural equivalencies between NMDS solutions and factor models. It could be shown that NMDS solutions constitute a promising framework to derive empirically based factor models. These findings are applicable to virtually any item structure and thus are not limited to depression or psychopathological inventories. Furthermore, the results link the rather unknown NMDS analysis with the renowned factor analysis, which simplifies the interpretation of NMDS solutions and their integration in previous findings. Hopefully, the manuscript might contribute to further spread NMDS analyses and their application in the examination of item structures.

Future directions

NMDS analyses of symptom profiles yield great potential to be of practical use, not only at admission (for example to help infer the diagnosis of a patient, as was shown by Egli et al. (2009), but also during the course of treatment. The continuous analysis of symptom profiles in patient maps provides immediate feedback to clinicians. Non response or symptomatic shifts during treatment can thus be promptly detected. Furthermore, similar patients are easily identifiable, which grants fast access to their treatment regimen and, granted their treatment has already finished, to their response to treatment. As a consequence, patient maps yield the potential to substantially boost evidence based medicine in psychiatry.

However, as was noted in the last manuscript, there are still many imponderables in the calculation of patient maps. Thus, future studies are needed to assess the patient maps' general applicability. Preliminary results on the stability of patient maps in depression (not discussed in

this thesis) suggest that random symptom profile samples might partition the patient map differently and thus highlight highly sample specific characteristics of a disorder. Thus, among the most pressing topics of future research in the development of patient maps for clinical practice resides the development and research of standardized patient maps (i.e. patient maps based on a representative set of symptom profiles) to increase stability of NMDS results. Furthermore, a systematic evaluation whether a two dimensional patient space is able to sufficiently structure the patient and disorder specific characteristics would certainly fortify the applicability of patient maps and broaden the knowledge on the examined disorders. However, chances are that the dimensionality needs to vary with the specific disorder category and/or psychopathological inventory.

With respect to the symptom maps, the benefits of NMDS analyses were demonstrated by integrating previous research efforts in a coherent model. The good comparability of factor models on the grounds of NMDS solutions may foster the NMDS's application in the analysis of item structures. Thus, future researchers are encouraged to apply NMDS to additional psychopathological inventories and explore the structure of mental disorders beyond depression.

The benefit of symptom maps in clinical practice is straight forward: individual patients' symptom profiles (e.g. a BDI-II profile) can be depicted graphically in the respective symptom map by color coding the symptoms included in the map. For example, the color may vary with the observed symptom score: yellow for low, orange for medium and red for high. Thus, the figure could deliver a rapid assessment of a patient's severity while preserving all the details (i.e. severity not only as a global, but as a measure on the level of symptoms). Furthermore, the symptoms' inherent correlational structure would be conveyed implicitly by their locations on the map.

The application of symptom and patient maps in clinical practice seem beneficial because clinicians rely heavily on symptom specific information when selecting treatments for their patients (Zimmerman et al., 2004). Furthermore, previous research suggests great benefit from feedback in psychotherapy (Slade et al., 2008). However, efficacy of these instruments has not yet been tested. Thus, further research is needed to prove an increase in efficacy by the use of symptom and patient maps in clinical practice.

Concluding remarks

The studies included in this thesis extended the previous findings by Egli et al. (Egli et al., 2009) and by Läge et al. (Damian Läge et al., 2012) with detailed insights into the structure of two of the most common rating scales of depression, the BDI-II and the HAM-D. For future research, a generally applicable outcome of this thesis is the extensively discussed close link between NMDS solutions and factor models, both of simple and complex factor structures. It

highlights a general principle of the items locations in NMDS solutions and the corresponding factor loadings. Thus, the results of factor analyses can be discussed on the basis of NMDS solutions, and, vice versa, new factor models can be derived from NMDS solutions. Furthermore, the proposed methodological advancement (bootstrap) in NMDS analyses allows estimating the precision of the items' locations in NMDS solutions via confidence regions. The calculated NMDS solutions can thus be evaluated with respect to their stability and precision. Additionally, an extended procedure of conducting NMDS analyses was evaluated, and it was shown to improve the items location estimates. With respect to the patient maps, the required research and methodological difficulties of applying NMDS in the course of treatment was emphasized. Lastly, empirical evidence was collected, which suggests that specific symptom patterns influence the response to treatment in depression. These findings were an essential step along the way towards an application of patient maps in clinical practice.

Hopefully, this thesis will stimulate future research in the field of psychopathology to apply NMDS analyses. In my personal opinion, this may lead the way to not only a better understanding of the structure of mental disorders, but also to increase evidence-based medicine in psychological and psychiatric care.

References

Abdi, H., Dunlop, J. P., & Williams, L. J. (2009). How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the Bootstrap and 3-way multidimensional scaling (DISTATIS). *NeuroImage*, *45*(1), 89-95.

Addington, D., Addington, J., & Atkinson, M. (1996). A psychometric comparison of the Calgary depression scale for schizophrenia and the Hamilton depression rating scale. *Schizophrenia Research*, *19*(2-3), 205-212.

Aggen, S. H., Neale, M. C., & Kendler, K. S. (2005). DSM criteria for major depression: evaluating symptom patterns using latent-trait item response models. *Psychological Medicine*, *35*(4), 475-487.

Akdemir, A., Türkçapar, M. H., Örsel, S. D., Demirergi, N., Dag, I., & Özbay, M. H. (2001). Reliability and validity of the Turkish version of the Hamilton Depression Rating Scale. *Comprehensive Psychiatry*, *42*(2), 161-165.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Arbeitsgemeinschaft für Methodik und Dokumentation in der Psychiatrie (1981). *Das AMDP-system. Manual zur Dokumentation psychiatrischer Befunde*. Berlin: Springer.

Arnau, R. C., Meagher, M. W., Norris, M. P., & Bramson, R. (2001). Psychometric evaluation of the Beck Depression Inventory-II with primary care medical patients. *Health Psychology 20*(2), 112-119.

Asparouhov, T., & Muthén, B. (2010). *Simple second order chi-square correction*. Technical appendix. Los Angeles: Muthén & Muthén. Retrieved November 2011, from https://www.statmodel.com/download/WLSMV_new_chi21.pdf

Bagby, R. M., Ryder, A. G., Schuller, D. R., & Marshall, M. B. (2004). The Hamilton Depression Rating Scale: Has the Gold Standard Become a Lead Weight? *American Journal of Psychiatry*, *161*(12), 2163-2177.

Bagby, R. M., Ryder, A. G., Schuller, D. R., & Marshall, M. B. (2004). The Hamilton Depression Rating Scale: has the gold standard become a lead weight? *The American journal of psychiatry*, *161*(12), 2163-2177.

Baumeister, H., & Parker, G. (2012). Meta-review of depressive subtyping models. *Journal of Affective Disorders*, *139*(2), 126-140.

Bech, P., Allerup, P., Gram, L. F., Reisby, N., Rosenberg, R., Jacobsen, O., & Nagy, A. (1981). The Hamilton Depression Scale. Evaluation of objectivity using logistic models. *Acta psychiatrica Scandinavica*, *63*(3), 290-299.

Bech, P., Gram, L. F., Dein, E., Jacobsen, O., Vitger, J., & Bolwig, T. G. (1975). Quantitative rating of depressive states. *Acta Psychiatrica Scandinavica, 51*(3), 161-170.

Beck, A. T., Steer, R. A. & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory – Second Edition. Manual.* San Antonio, TX: The Psychological Corporation.

Beck, A. T., Ward, C. H., Mendelsohn, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of general Psychiatry 4*(6), 561-571.

Bedi, R. P., Koopman, R. F. & Thompson, J. M. (2001). The dimensionality of the Beck Depression Inventory-II and its relevance for tailoring the psychological treatment of women with depression. *Psychotherapy: Theory, Research, Practice, Training, 38*, 306-318.

Blazer, D., Woodbury, M., Hughes, D. C., George, L. K., Manton, K. G., Bachar J. R., & Fowler, N. (1989). A statistical analysis of the classification of depression in a mixed community and clinical sample. *Journal of Affective Disorders, 16*, 11-20.

Borg, I., & Groenen, P. J. (2005). *Modern Multidimensional Scaling, theory and applications* (2nd ed.). New York: Springer.

Borg, I., & Staufenbiel, T., (2007*). Lehrbuch - Theorien und Methoden der Skalierung* (4th ed.). [Textbook – Theory and Methods of Scaling. (4th edn.)]. Hans Huber AG: Bern.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). On the factor structure of the Beck Depression Inventory-II: G is the key. *Psychological Assessment 25*(1), 136-145.

Brown, C., Schulberg, H. C., & Madonia, M. J. (1995). Assessing depression in primary care practice with the Beck Depression Inventory and the Hamilton Rating Scale for Depression. *Psychological Assessment, 7*(1), 59-65.

Brown, T. A., & Barlow, D. (2009). A Proposal for a Dimensional Classification System Based on the Shared Features of the DSM-IV Anxiety and Mood Disorders: Implications for Assessment and Treatment. *Psychological Assessment, 21*(3), 256-271.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of model fit. In K. A. Bollen, & J. S. Long, *Testing structural equation models* (136 - 162). Newbury Park, CA: Sage.

Buckley, T. C., Parker, J. D., & Heggie, J. (2001). A psychometric evaluation of the BDI II in treatment-seeking substance abusers. *Journal of Substance Abuse Treatment 20*(3), 197-204.

Bühler, J., Läge, D. (2013). *Better bootstrap NMDS analyses – confidence regions and improved location estimates in Nonmetric Multidimensional Scaling*; Manuscript in preparation.

Bühler, J., Keller, F., & Läge, D. (2012). Die Symptomstruktur des BDI-II: Kernsymptome und qualitative Facetten. [The symptom structure of the BDI-II: core symptoms and qualitative facets]. *Zeitschrift für Klinische Psychologie und Psychotherapie, 41*(4), 231-242.

Bühler, J., Keller, F., & Läge, D. (2013). *Activation as an overlooked factor in the BDI-II: a factor model based on core symptoms and qualitative aspects of depression.* Manuscript submitted for publication.

Bühler, J., Seemüller F., & Läge, D. (2013). *The Symptom Structure of the Hamilton Depression Scale (HAM-D): an NMDS analysis*. Manuscript submitted for publication.

Carragher, N., Adamson, G., Bunting, B., & McCann, S. (2009). Subtypes of depression in a nationally representative sample. *Journal of Affective Disorders, 113*(1) 88-99.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35*(3), 283-319.

Carroll, J. D., & Green, P. E. (1997). Psychometric methods in marketing research: Part II, multidimensional scaling. *Journal of Marketing Research, 34,* 193-204.

Chen, L., Eaton, W., Gallo, J., & Nestadt, G. (2000). Understanding the heterogeneity of depression through the triad of symptoms, course and risk factors: a longitudinal, population-based study. *Journal of Affective Disorders, 59*(1), 1-11.

Cohen, A. (2008). The underlying structure of the Beck Depression Inventory II: a Multidimensional Scaling approach. *Journal of Research in Personality, 42*(3), 779-786.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, J. C., Motivala, S. J., Dang, J., Lucko, A., Lang, N., Levin, M. J., … Irwin, M. (2004). Structural validation of the Hamilton Depression Rating Scale. *Journal of Psychopathology and Behavioral Assessment, 26*(4), 241-254.

Collegium Internationale Pyschiatriae Scalarum (CIPS). (1977). *Internationale Skalen für Psychiatrie* [International scales for psychiatry]. Weinheim: Beltz Test.

Commandeur, J. J. F., & Heiser, W. J. (1993). *Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices* (Tech. Rep. No. RR-93-03). Leiden, The Netherlands: Department of Data Theory, Leiden University.

Cox, B. J., Enns, M. W., & Larsen, D. K. (2001). The continuity of depression symptoms: Use of cluster analysis for profile identification in patient and student samples. *Journal of Affective Disorders, 65,* 67-73.

Cox, M. A. A. (2012). Analysis of stock market indices through multidimensional scaling. *Journal of Statistical Computation and Simulation*. Retrieved from http://dx.doi.org/10.1080/00949655.2012.678361

Cuijpers, P., Straten, A., Bohlmeijer, E., Hollon, S. D., Andersson, G., & van Straten, A. (2010). The effects of psychotherapy for adult depression are overestimated: a meta-analysis of study quality and effect size. *Psychological Medicine, 40*(2), 211-223.

Damm, J., Eser, D., Schüle, C., Möller, H. -J., Rupprecht, R., & Baghai, T. C. (2009). Depressive Kernsymptome. [Depressive Core Symptoms]. *Der Nervenarzt, 80*(5), 515-531.

De Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In J. R. Barra, F. Brodeau, G. Romier, & B. van Cutsem (Eds.), *Recent developments in statistics* (pp. 133-145). Amsterdam, The Netherlands: North-Holland.

Dozois, D. J., Dobson, K. S. & Ahnberg, J. L. (1998). A psychometric evaluation of the Beck Depression Inventory-II. *Psychological Assessment, 10*, 83-89.

Driessen, E., & Hollon, S. (2010). Cognitive Behavioral Therapy for mood disorders: efficacy, moderators and mediators. *Psychiatric Clinics of North America*, *33*(3), 537-555.

Eaton, W., Dryman, A., Sorenson, A., & McCutcheon, A. (1989). DSM-III major depressive disorder in the community. A latent class analysis of data from the NIMH epidemiologic catchment area programme. *The British Journal of Psychiatry*, *155*(1), 48-54.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics, 7*(1), 1-26.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1*(1), 54-77.

Egli, S., Riedel, M., Möller, H., Strauss, A., & Läge, D. (2009). Creating a map of psychiatric patients based on psychopathological symptom profiles. *European Archives of Psychiatry and Clinical Neuroscience, 259*(3), 164-171.

Esposito, K., & Goodnick, P. (2003). Predictors of response in depression. *Psychiatric Clinics of North America*, *26*(2), 353-365.

Evans, K. R., Sills, T., DeBrota, D. J., Gelwicks, S., Engelhardt, N., & Santor, D. (2004). An Item Response analysis of the Hamilton Depression Rating Scale using shared data from two pharmaceutical companies. *Journal of Psychiatric Research*, *38*(3), 275-284.

Faries, D., Herrera, J., Rayamajhi, J., DeBrota, D., Demitrack, M., & Potter, W. Z. (2000). The responsiveness of the Hamilton Depression Rating Scale. *Journal of Psychiatric Research*, *34*(1), 3-10.

Fava, M., Uebelacker, L., Alpert, J., Nierenberg, A., Pava, J., & Rosenbaum, J. (1997). Major depressive subtypes and treatment response. *Biological Psychiatry*, *42*(7), 568-576.

Fink, M., & Taylor, M. A. (2007). Resurrecting melancholia. *Acta psychiatrica Scandinavica, 115*(Suppl. 433), 14-20.

Fleck, M., Poirier-Littre, M., & Guelfi, J. (1995). Factorial structure of the 17-item Hamilton Depression Rating Scale. *Acta Psychiatrica Scandinavica, 92*, 168-172.

Fournier, J. C., DeRubeis, R. J., Shelton, R. C., Hollon, S. D., Amsterdam, J. D., & Gallop, R. (2009). Prediction of response to medication and cognitive therapy in the treatment of moderate to severe depression. *Journal of Consulting and Clinical Psychology*, *77*(4), 775-787.

Fraley, C., & Raftery, A. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, *97*(458), 611-631.

Fraley, C., Raftery, A., Murphy, T., & Scrucca, L. (2012). *mclust Version 4 for R: Normal Mixture Modeling for model-based clustering, classification, and density estimation.* (Technical Report No. 597, Departement of Statistics, University of Washington). Retrieved from http://www.stat.washington.edu/research/reports/2012/tr597.pdf

Frick, P. J., Lahey, B. B., Loeber, R., Tannenbaum, L., van Horn, Y., Christ, M. A., … Hanson, K. (1993). Oppositional defiant disorder and conduct disorder: A meta-analytic review of factor analyses and cross-validation in a clinic sample. *Clinical Psychology Review, 13*(4), 319-340.

Gebhardt, R., Pietzcker, A., Strauss, A., Stöckel, M., Langer, C., & Freudenthal, K. (1983). Skalenbildung im AMDP-System. *Archiv für Psychiatrie und Nervenkrankheiten, 233*, 223-245.

Gibbons, R. D., Clark, D. C., & Kupfer, D. J. (1993). Exactly what does the Hamilton depression rating scale measure? *Journal of Psychiatric Research, 27*(3), 259-273.

Gigerenzer, G. (1981). *Messung und Modellbildung in der Psychologie.* München, Basel: E. Reinhardt.

Gower, J. C., & Dijksterhuis, G. B. (2004). *Procrustes Problems*. Oxford: University Press.

Groenen, P. J., & Heiser, W. J. (1996). The tunneling method for global optimization in multidimensional scaling. *Psychometrika, 61*(3), 529-550.

Hamdi, E., Amin, Y., & Abou-Saleh, M. T. (1997). Performance of the Hamilton Depression Rating Scale in depressed patients in the United Arab Emirates. *Acta Psychiatrica Scandinavica, 96*(6), 416-423.

Hamilton, K. E., & Dobson, K. S. (2002). Cognitive therapy of depression: pretreatment patient predictors of outcome. *Clinical Psychology Review, 22*, 875-893.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23*(1), 56-62.

Hammond, M. (1998). Rating depression severity in the elderly physically ill patient: reliability and factor structure of the Hamilton and the Montgomery-Asberg Depression Rating Scales. *International Journal of Geriatric Psychiatry, 13*(4), 257-261.

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of clinical psychology, 61*(2), 155-163.

Hatfield, D. R., & Ogles, B. M. (2006). The influence of outcome measures in assessing client change and treatment decisions. *Journal of Clinical Psychology, 62*(3), 325-337.

Haughton, D., Legrand, P., & Woolford, S. (2009). Review of three latent class cluster analysis packages: Latent Gold, poLCA, and MCLUST. *The American Statistician, 63*(1), 81-91.

Hautzinger, M., Bailer, M., Worall, H., & Keller, F. (1995). *Beck-Depressions-Inventar (BDI). Testhandbuch (2. Auflage).* [Beck Depression Inventory. Manual (2nd ed.)]. Bern: Huber.

Hautzinger, M., Keller, F., & Kühner, C. (2006). *BDI-II. Beck Depressions Inventar Revision – Manual.* [BDI-II. Revision of the Beck Depression Inventory – Manual]. Frankfurt: Harcourt Test Services.

Heiser, W. J., & Meulman, J. (1983). Constrained Multidimensional Scaling, Including Confirmation. *Applied Psychological Measurement, 7*(4), 381-404.

Helmreich, I., Wagner, S., Mergl, R., Allgaier, A., Hautzinger, M., Henkel, V., … Tadić, A. (2012). Sensitivity to changes during antidepressant treatment: a comparison of unidimensional subscales of the Inventory of Depressive Symptomatology (IDS-C) and the Hamilton Depression Rating Scale (HAMD) in patients with mild major, minor or subsyndromal depression. *European archives of psychiatry and clinical neuroscience, 262*(4), 291-304.

Hollon, S. D., & Najavits, L. (1988). Review of empirical studies on cognitive therapy. In A. Frances & R. Hales (Eds.), *Psychiatry update: American Psychiatric Association annual review* (Vol. 7, pp. 643-666). Washington DC: American Psychiatry Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal 6*(1), 1-55.

Jarrett, R. B., Minhajuddin, A., Kangas, J. L., Friedman, E. S., Callan, J. A., & Thase, M. E. (2013). Acute phase cognitive therapy for recurrent major depressive disorder: Who drops out and how much do patient skills influence response. *Behaviour research and therapy, 51*(4-5), 221-230.

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika, 76*(4), 537-549.

Joiner, T. E., Walker, R. L., Pettit, J. W., Perez, M., & Cukrowicz, K. C. (2005). Evidence-based assessment of depression in adults. *Psychological Assessment*, 17(3), 267-277.

Keller, F., Hautzinger, M., & Kühner, C. (2008). Zur faktoriellen Struktur des deutschsprachigen BDI-II. [Factor structure of the German Beck Depression Inventory, Second Edition (BDI-II)]. *Zeitschrift für Klinische Psychologie und Psychotherapie, 37*(4), 245-254.

Kendler, K. S., Eaves, L. J., Walters, E. E., Neale, M. C., Heath, A. C., & Kessler, R. C. (1996). The identification and validation of distinct depressive syndromes in a population-based sample of female twins. *Archives of General Psychiatry, 53*(5), 391-399.

Klein, D. F., Davis J. M. (1969). *Diagnosis and drug treatment of psychiatric disorders*. Baltimore: Williams & Wilkins.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Koukopoulos, A., & Koukopoulos, A. (1999). Agitated depression as a mixed state and the problem of melancholia. *The Psychiatric clinics of North America, 22*(3), 547-64.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*(1), 1-27.

Kruskal, J. B. (1964b). Nonmetric Multidimensional Scaling: a Numerical Method. *Psychometrika, 29*(2), 115-129.

Kühner, C., Bürger, C., Keller, F. & Hautzinger, M. (2007). Reliabilität und Validität des revidierten Beck Depressionsinventars (BDI-II) - Befunde aus deutschsprachigen Stichproben. *Der Nervenarzt, 78*, 651-656.

Läge, D., Daub, S., Bosia, L., Jäger, C., & Ryf, S. (2005). Die Behandlung ausreisserbehafteter Datensätze in der Nonmetrischen Multidimensionalen Skalierung - Relevanz, Problemanalyse und Lösungsvorschlag. [Handling of outlier afflicted datasets in Nonmetric Multidimensional Scaling - relevance, problem analysis and approach for a solution]. *Forschungsberichte aus der Angewandten Kognitionspsychologie Zürich*, 1-33.

Läge, D., Egli, S., Riedel, M., & Möller, H. J. (2012). Exploring the structure of psychopathological symptoms: a re-analysis of AMDP data by robust nonmetric multidimensional scaling. *European Archives of Psychiatry and Clinical Neuroscience, 262*(3), 227-238.

Läge, D., Egli, S., Riedel, M., Strauss, A., & Möller, H. (2011). Combining the categorical and the dimensional perspective in a diagnostic map of psychotic disorders. *European Archives of Psychiatry and Clinical Neuroscience*, *261*(1), 3-10.

Lambert, M., University, B., University, P., Harmon, C., Slade, K., Whipple, J., u. a (2005). Providing feedback to psychotherapists on their patients' progress: Clinical results and practice suggestions. *Journal of Clinical Psychology*, *61*(2), 165-174. doi:10.1002/jclp.20113

Lichtenberg, P., & Belmaker, R. (2010). Subtyping Major Depressive Disorder. *Psychotherapy and Psychosomatics*, *79*(3), 131-135.

Lingoes, J. C. & Roskam, E. E. (1973). *A Mathematical and Empirical Analysis of Two Multidimensional Scaling Algorithms* (Psychometric Monograph No. 19). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN19.pdf

Loeber, R., & Schmaling, K. B. (1985). Empirical evidence for overt and covert patterns of antisocial conduct problems: A metaanalysis. *Journal of Abnormal Child Psychology, 13*(2), 337-353.

Maier, W., & Philipp, M. (1985). Comparative analysis of observer depression scales. *Acta Psychiatrica Scandinavica*, *72*(3), 239-245.

Marcos, T., & Salamero, M. (1990). Factor study of the Hamilton Rating Scale for Depression and the Bech Melancholia Scale. *Acta Psychiatrica Scandinavica*, *82*(2), 178-181.

Mathar, R., & Groenen, P. J. F. (1991). Algorithms in convex analysis applied to multidimensional scaling. In E. Diday & Y. Lechevallier (Eds.), *Symbolic-numeric data analysis and learning*, (pp. 45-56). Commack, NY: Nova Science Publishers.

McIntyre, R., Kennedy, S., Bagby, R. M., & Bakish, D. (2002). Assessing full remission. *Journal of Psychiatry and Neuroscience*, *27*(4), 235-239.

Michaux, M., Suziedelis, A., Garmize, K., Suziedelis, A., Garmize, K., Suziedelis, A., & Garmize, K. (1969). Depression factors in depressed and in heterogeneous in-patient samples. *Journal of Neurology, Neurosurgery and Psychiatry*, *32*(6), 609-613.

Möller, H. (2001). Methodological aspects in the assessment of severity of depression by the Hamilton Depression Scale. *European Archives of Psychiatry and Clinical Neuroscience*, *251*(Suppl. 2), 13-20.

Möller, H. J. (2009). Standardised rating scales in psychiatry: methodological basis, their possibilities and limitations and descriptions of important rating scales. *World Journal of Biological Psychiatry*, *10*(1), 6-26.

Montani, T. (2009). *Die Zürcher Stufenplanstudie: Vergleich systematischer Therapiealgorithmen mit der Standardbehandlung bei Patienten mit unipolarer Depression und deren neuropsychologische Charakterisierung im Verlauf* (Unpublizierte Doktorarbeit). [The Zurich stepwise treatment plan: A comparison of systematic treatment algorithms with treatment as usual in a sample of patients with unipolar depression and their neuropsychological characteristics over time (unpublished doctoral thesis)]. University of Zurich.

Moran, P. W. & Lambert, M. J. (1983). A review of current assessment tools for monitoring changes in depression. In M. S. Lambert, E. R. Christensen, & S. DeJulio, *The assessment of psychotherapy outcome* (S. 263-303). New York: Wiley.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide* (7nth ed.). Los Angeles, CA: Muthén & Muthén.

Oberholzer, R., Egloff, S., Ryf, S., & Läge, D. (2008). *Prodax Handbuch*. Zürich: Authors. Retrieved from http://www.prodax.ch/ProDaX-Handbuch.pdf.

O'Brien, K. P., & Glaudin, V. (1988). Factorial structure and factor reliability of the Hamilton Rating Scale for Depression. *Acta Psychiatrica Scandinavica*, *78*(2), 113-120.

Onega, L., & Abraham, I. (1997). Factor structure of the Hamilton Rating Scale for Depression in a cohort of community dwelling elderly. *International Journal of Geriatric Psychiatry*, *12*(7), 760-764.

Osman, A., Downs, W. R., Barrios, F. X., Kopper, B. A., Guttierrez, P. M., & Chiros, C. E. (1997). Factor structure and psychometric characteristics of the Beck Depression Inventory-II. *Journal of Psychopathology and Behavioral Assessment 19*(4), 359-376.

Osman, A., Kopper, B. A., Barrios, F., Gutierrez, P. M. & Bagge, C. L. (2004). Reliability and validity of the Beck depression inventory-II with adolescent psychiatric inpatients. *Psychological Assessment, 16*, 120-132.

Padilla, J.-L., Benitez, I., Sireci, S. G., & Flores-Galaz, M. (2012). Evaluating Structural Equivalence in Psychological Questionnaires Using Weighted Multidimensional Scaling. *Cross-Cultural Research, 46*(4), 348-365.

Pancheri, P., Picardi, A., Pasquini, M., Gaetano, P., & Biondi, M. (2002). Psychopathological dimensions of depression: a factor study of the 17-item Hamilton Depression Rating Scale in unipolar depressed outpatients. *Journal of Affective Disorders, 68*(1), 41-47.

Parker, G. (2007). Defining melancholia: the primacy of psychomotor disturbance. *Acta Psychiatrica Scandinavica, 115*(Suppl. 433), 21-30.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Development Core Team (2013). *nlme: Linear and Nonlinear Mixed Effects Models* (R package version. 3.1-105).

Qin, S., Wan, Q., & Duan, L.-F. (2012). Fast and efficient multidimensional scaling algorithm for mobile positioning. *IET Signal Processing, 6*(9), 857-861.

Quilty, L. C., Zhang, K. A., & Bagby, R. M. (2010). The latent symptom structure of the Beck Depression Inventory-II in outpatients with major depression. *Psychological Assessment, 22*(3), 603-608.

R Development Core Team. (2012). *R: A Language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Ramos-Brieva, J. A., & Cordero-Villafafila, A. (1988). A new validation of the Hamilton Rating Scale for depression. *Journal of Psychiatric Research, 22*(1), 21-28.

Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika, 42*(2), 241-266.

Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling data. *Journal of the Royal Statistical Society. Series A, 145*(3), 285-312.

Rao, S., Zisook, S. (2009). Anxious depression: clinical features and treatment. *Current Psychiatry Reports, 11,* 429-436.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reynolds, W. M., & Kobak, K. A. (1995). Reliability and validity of the Hamilton Depression Inventory: A paper-and-pencil version of the Hamilton Depression Rating Scale Clinical Interview. *Psychological Assessment, 7*(4), 472-483.

Santen, G., Gomeni, R., Danhof, M., & della Pasqua, O. (2008). Sensitivity of the individual items of the Hamilton depression rating scale to response and its consequences for the assessment of efficacy. *Journal of Psychiatric Research, 42*(12), 1000-1009.

Santor, D. A., & Coyne, J. C. (2001). Examining symptom expression as a function of symptom severity: Item performance on the Hamilton Rating Scale for Depression. *Psychological Assessment, 13*(1), 127-139.

Santor, D. A., Gregus, M. & Welch, A. (2006). Eight decades of measurement in depression. *Measurement, 4*(3), 135-155.

Schmid, J. & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53-61.

Shafer, A. B. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology, 62*(1), 123-146.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27*(2), 125-140.

Shorter, E. (2007). The doctrine of the two depressions in historical perspective. *Acta Psychiatrica Scandinavica, 115*(Suppl. 433), 5-13.

Slade, K., Lambert, M., Harmon, S. C., Smart, D. W., & Bailey, R. (2008). Improving psychotherapy outcome: the use of immediate electronic feedback and revised clinical support tools. *Clinical Psychology & Psychotherapy, 15*(5), 287-303.

Spence, I., & Lewandowsky, S. (1989). Robust multidimensional scaling. *Psychometrika, 54*(3), 501-513.

Spitzer, R. L., Endicott, J., & Robins, E. (1978). Research Diagnostic Criteria: rationale and reliability. *Archives of General Psychiatry, 35*, 773-782.

Steer, R. A. & Clark, D. A. (1997). Psychometric characteristics of the Beck Depression Inventory-II with college students. *Measurement & Evaluation in Counseling & Development, 30*, 128-136.

Steer, R. A., Ball, R., Ranieri, W. F., & Beck, A. T. (1999). Dimensions of the Beck Depression Inventory-II in clinically depressed outpatients. *Journal of Clinical Psychology, 55*(1), 117-128.

Steinmeyer, E. M., & Möller, H.-J. (1992). Facet theoretic analysis of the Hamilton-D scale. *Journal of Affective Disorders, 25*(1), 53-62.

Steinmeyer, E.M. (1993). Zur klinischen Validität des Beck Depressionsinventars: Eine facettentheoretische Reanalyse multizentrischer klinischer Beobachtungsdaten. [About the clinical validity of the Beck Depression Inventory: a facet-theoretical re-analysis of clinical multicenter data]. *Der Nervenarzt, 64*, 717-726.

Stewart, J. W., Garfinkel , R., Nunes, E. V., Donovan, S. & Klein, D. F. (1998). Atypical features and treatment response in the National Institute of Mental Health Treatment of Depression Collaborative Research program. *Journal of Clinical Psychopharmacology, 18*(6), 429-434.

Stewart, J. W., McGrath, P. J., Quitkin, F. M., & Klein, D. F. (2007). Atypical depression: current status and relevance to melancholia. *Acta psychiatrica Scandinavica, 115*(Suppl. 433), 58-71.

Sturrock, K., & Rocha, J. (2000). A Multidimensional Scaling Stress Evaluation Table. *Field Methods, 12*(1), 49-60.

Sullivan, P. F., Prescott, C. A., & Kendler, K. S. (2002). The subtypes of major depression in a twin registry. *Journal of Affective Disorders, 68*(2-3), 273-84.

Tang, T. Z., & DeRubeis, R. J. (1999). Sudden Gains and Critical Sessions in Cognitive-Behavioral Therapy for Depression. *Journal of Consulting and Clinical Psychology, 67*(6), 894-904.

Thurstone, L. L. (1954). An analytical method for simple structure, *Psychometrika, 19*(3), 173-182.

Torgerson, W. S. (1952). Multidimensional Scaling: I. Theory and method. *Psychometrika, 17*(4), 401-419.

Torgerson, W. S. (1958). *Theory and Methods of Scaling*. New York: Wiley.

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *The New England Journal of Medicine, 358*(3), 252-260.

Vanheule, S., Desmet, M., Groenvynck, H., Rosseel, Y., & Fontaine, J. (2008). The factor structure of the Beck Depression Inventory-II: an evaluation. *Assessment 15*(2), 177-187.

Vanpoucke, F. J., Boermans, P. B., & Frijns, J. H. (2012). Assessing the Placement of a Cochlear Electrode Array by Multidimensional Scaling. *IEEE Transactions on Biomedical Engineering, 59*(2), 307-310.

Viljoen, J. L., Grant, L. I., Griffiths, S., & Woodward, T. S. (2003). Factor structure of the Beck Depression Inventory-II in a medical outpatient sample. *Journal of Clinical Psychology in Medical Settings, 10*(4), 289-291.

von Giesen, H. J., Bäcker, R., Hefter, H., & Arendt, G. (2001) Depression does not influence basal ganglia-mediated psychomotor speed in HIV-1 infection. *The Journal of neuropsychiatry and clinical neurosciences, 13*(1), 88-94.

von Oertzen, T., Brandmaier, A. M., & Tsang, S. (2012). *The Onyx user guide V0.1*. Retrieved December 2012, from http://onyx.brandmaier.de/userguide.pdf.

Ward, L. C. (2006). Comparison of factor structure models for the Beck Depression Inventory-II. *Psychological Assessment 18*(1), 81-88.

Weckowicz, T., Cropley, A., & Muir, W. (1971). An attempt to replicate the results of a factor analytic study in depressed patients. *Journal of Clinical Psychology, 27*(1), 30-31.

Weinberg, S. L., Carroll, J. D., & Cohen, H. S. (1984). Confidence regions for INDSCAL using the jackknife and bootstrap techniques. *Psychometrika, 49*(4), 475-491.

Whisman, M. A., Perez, J. E., & Ramel, W. (2000). Factor structure of the Beck Depression Inventory-Second Edition (BDI-II) in a student sample. *Journal of Clinical Psychology 56*(4), 545-551.

Williams, J. B., Kobak, K. A., Bech, P., Engelhardt, N., Evans, K., Lipsitz, J., … Kalali, A. (2008). The GRID-HAMD: standardization of the Hamilton depression rating scale. *International Clinical Psychopharmacology, 23*(3), 120.

Wittchen, H. U., Wunderlich, U., Gruschwitz, S., & Zaudig, M. (1997). *Strukturiertes Klinisches Interview für DSM-IV* [Structured clinical interview for DSM-IV]. Göttingen: Hogrefe.

World Health Organization (1992). *The ICD-10 Classification of Mental and Behavioral Disorders*. Geneva: World Health Organization

Zimmerman, M., Posternak, M. A., & Chelminski, I. (2005). Is it time to replace the Hamilton Depression Rating Scale as the primary outcome measure in treatment studies of depression? *Journal of Clinical Psychopharmacology, 25*(2), 105-110.

Zimmerman, M., Posternak, M., Friedman, M., Attiullah, N., Baymiller, S., Boland, R., … Singer, S. (2004). Which factors influence psychiatrists' selection of antidepressants? *The American Journal of Psychiatry, 161*(7), 1285-1289.

Acknowledgements

First and foremost I would like to express my sincere gratitude to my supervisor Prof. Dr. Damian Läge for his continuous support and feedback throughout my doctorate. His concise reasoning and his courage to think outside the box has always been a great source of inspiration.

My thanks go to Prof. Dr. Ferdinand Keller at the University of Ulm, with whom I had the pleasure to write the manuscripts on the BDI-II. His profound knowledge of MPlus and his critical thoughts greatly contributed to both manuscripts.

I would also like to thank Prof. Dr. Carolin Strobl, who reviewed this thesis as a member of the supervisory committee, and who provided me with many useful hints regarding the most methodological paper on the bootstrap in NMDS.

My sincere appreciation goes to Lars von Mühlenen, without whom the funding of this project would not have been possible. I wish him all the best in promoting the PELION services and thus, hopefully, in improving the handling of psychopathological inventories in clinical practice.

I am deeply indebted to my girlfriend Stefanie Huber, who encouraged me in times of motivational setbacks and whose mere presence was an essential source of comfort and calm during the more stressful periods of writing this thesis.

Last, but certainly not least, I would like to thank my family, Silvie, Jean-Marc and Nathalie, whose support and encouragement have always kept me on track to keep pursuing my dreams.

Curriculum Vitae
Joël Bühler

personal information

address	Färberstrasse 10 8400 Zürich Switzerland
e-mail	joel.buehler@uzh.ch
phone	+41 79 662 63 22
date of birth	31 January 1983
nationality	Swiss
marital status	unmarried

education & qualifications

03/2011-10/2013 PhD student at the departement of Psychology, University of Zurich
- applied cognitive psychology (modelling of the symptom structure of depression and individual depressive patients' profiles)

10/2003–10/2010 M. Sc. Psychology, University of Zurich
- major: General Psychology (Cognition)
- minor (60 CP): Physics
- thesis: Psychoacoustics

work history

03/2011-10/2013 PhD-student at the University of Zurich
- focus on the application and enhancement of quantitative methods. Wide range of analysis methods (NMDS, CFA, IRT, GLM) to examine psychological data structures.
- development supervision of a web-based tool to bring the scientific findings to the market (i.e. to psychiatric hospitals).

12/2010–02/2011 scientific assistant at the University of Zurich (35%):
- development of a set of methods to guide diagnostics, treatment and evaluation in psychiatric hospitals.

11/2009–09/2010 scientific assistant at the University of Zurich (50%):
- participation in a feasibility study on DRGs (Diagnostic Related Groups, funded by the Healthcare-Administration of the department of Zurich).
- development of a cost-per-patient allowance for psychiatric inpatients.

additional information

voluntary work & training
active-membership in PsyCH (umbrella organisation of psychology students) in promotion (2008)
active-membership in the methodology and statistics peer-mentoring group of the University of Zurich (since 2011)

November 2013